TechSonar

2025 REPORT

EDPS

European Data Protection Supervisor

Project Number: 2024.2409
Title: TechSonar reports

PDF/Volume_01	QT-AE-24-001-EN-N	978-92-9242-886-0	10.2804/797140

Exploring the future: Assessing the impact of emerging AI technologies on individuals

By Wojciech Wiewiórowski

2024 marks the 20th anniversary of the EDPS and the year the European Union adopts the Artificial Intelligence Act. Over the past two decades, the EDPS has played an important role in personal data protection at European level. We will continue to ensure that EU institutions, bodies, offices and agencies (EUIs) respect data protection rules when processing personal data. However, the role of the EDPS has recently expanded with the entry into force of the AI Act on 2 August 2024: we will act as the competent authority for AI systems provided and deployed by EUIs, ensuring a high level of protection against the potentially harmful effects of AI systems.

Many AI systems process personal data during their lifecycle (e.g. training, development or deployment). Consequently, artificial intelligence related technology trends are relevant for our two roles, as the AI Act's competent authority and as a data protection authority. With this in mind, we have decided to dedicate this issue of TechSonar entirely to AI technology trends, focusing on how these trends could impact the rights and freedoms of individuals.

This year's TechSonar report includes six trends: **Retrieval-augmented generation (RAG)**, a technique that allows AI systems to generate more relevant output by retrieving and combining relevant information from multiple knowledge bases. **On-device AI**, a system architecture designed to place data processing at the edge of the network, reducing latency and increasing control over the data processed by AI systems.

Machine unlearning, a technique that enables trained AI systems to forget specific data or remove its influence upon request. **Multimodal AI**, which deals with the integration of multiple types of data (e.g. text, images or audio), offering richer insights. **Scalable oversight**, focusing on the ability to use AI systems to effectively monitor other AI systems as they grow in complexity and scale, ensuring that AI applications remain transparent, accountable, and aligned with ethical standards. And finally, **neuro-symbolic AI**, which combines neural networks with symbolic reasoning to enhance accuracy and decision-making processes.

Each of these trends starts with **a fictional scenario** illustrating a potential application of the technology in our daily lives. Following the scenario, there is a description of the technology trend and its current development status. Then you will find our assessment of how the trend could affect individuals. To conclude each trend, we have compiled a list of recommended reading material for those wishing to gain a deeper understanding on the subject.

Without revealing too much about the trend reports ahead, I would like to share a few common elements and interesting patterns.

First, independently of the trend we consider, I observe that **most use cases of impactful AI applications process personal data.** It can easily be concluded that the deployment of AI systems in our daily lives will significantly rely on the processing of personal data. During the AI development and training phases, vast amounts of personal data, including text, images, audio, and video - often containing sensitive information such as biometric and behavioural data - are collected, posing significant risks such as potential data breaches, misuse or the incorporation of biased or unrepresentative data into AI models. Once trained, AI models may also memorise parts of their training datasets and be subject to data extraction attacks. Additionally, during the AI system deployment phase, user interactions with the models may involve further processing of personal data, raising privacy concerns, especially when biometric data is involved.

Second, some of these TechSonar **trends address challenges created by the way AI systems are currently developed.** For example, machine unlearning is linked to the problem of AI systems trained on poorly curated datasets, while retrieval-augmented generation contributes to solving the well-known problem of large language model hallucinations and scalable oversight relates to the cost of ensuring that increasingly complex AI systems are aligned with human values. Any technology that helps mitigate risks to human rights is welcome, but I wonder if we should first focus on avoiding the creation of these risks.

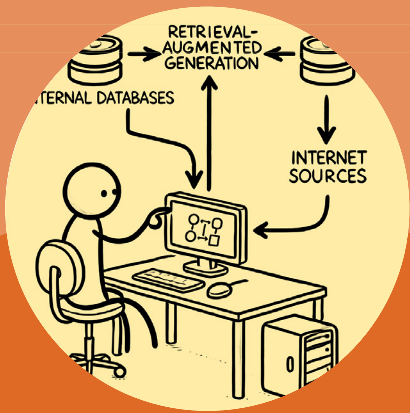
Third, while leading AI companies are introducing new AI models and trends at breath-taking speed, we see a **growing pressure on organizations to rapidly adopt and develop new AI systems quickly**. Given the new capabilities these new models and technology trends bring, there is growing fear among managers and employees within organizations of missing out on opportunities. These fears are understandable, but it is still paramount not to be dragged down by them as they could lead to poor risk management for individuals.

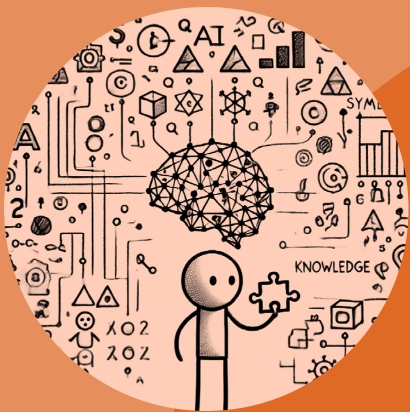
While some of the technologies in this TechSonar edition may contribute to mitigating the risks to individuals' fundamental rights, some, while promising substantial economic benefits, may also pose **significant risks to individuals if not properly managed**. The rapid progress in this area, combined with the potential for high returns on investment, is fuelling an AI race that is likely to make AI systems increasingly pervasive in our daily lives. This ubiquity of AI systems, combined with advanced capabilities that can profoundly influence our integrity and autonomy as persons, such as emotion detection, persuasive promotion of ideas, and content creation (including fake content), calls for certain limitations. Additional controls and safeguards are essential to ensure that the benefits of these advancements do not come at the expense of individual rights. Rigorous risk analysis and the implementation of robust safeguards are crucial. Providers and deployers of AI systems must conduct thorough impact assessments to identify potential risks and establish measures to mitigate them. This includes ensuring data privacy, preventing bias and discrimination, and maintaining transparency in AI operations. By prioritizing these safeguards, we can grab the full potential of AI technologies while preserving the fundamental rights of individuals.

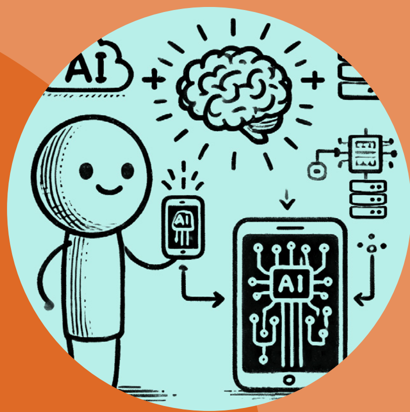
I hope this TechSonar can help disseminate knowledge about the key trends we see driving the field of AI for the coming years, and contribute to the ongoing debate by shedding light on the potential impacts - both positive and negative - of the AI trends presented. As with any foresight exercise, only time will reveal whether the technologies discussed here will evolve into major trends.

After all, reality is often more surprising than fiction!

Tech trends 2025







Note on the fictional scenarios provided

The EDPS provides fictional scenarios for each of the six AI-related trends.

It should be noted that the EDPS does not endorse these use cases.

Their purpose is merely to illustrate potential scenarios that could arise from the use of these technologies. They may include broad considerations relating to the fundamental rights to privacy and protection of personal data, as well as to society as a whole.

SCENARIO #1

A turn in the right direction...

A car rental company has recently deployed a new AI-based chatbot for customer support service. After several years in business, it became clear that effective 24/7 customer support was critical to the success of the business, as customers often contacted the company outside of office hours with issues that required an immediate response (e.g. how to report an accident or requests for roadside assistance).

Previously, telephone operators handled these requests, but during periods of high demand, such as the summer and Christmas holidays, there were often complaints about long waiting queues for operators to respond. This led the company to test an off-the-shelf chatbot system that proved to be only marginally better than the previous setup: the chatbot responses lacked insight on the business particularities and most times the chatbot directed the call to an operator.

Eventually, the company moved to another AI-based chatbot solution. The model was trained using documents collected during the company's operations over the years. This included information from previous customer incidents involving rental cars (such as accident reports and documents required by local authorities when crossing borders) and the corresponding steps the company took to resolve them, as well as previous customer complaints and frequently asked questions.

The new system allowed customers to ask questions and follow-up on the answers they were given to get more details on the issues at hand. By having access to a current list of partners (such as towing, car repair and legal services), they were able to provide customers in need with accurate contact details and opening hours.

Two months after deploying the new system the company made a survey among its customers and identified an overall increase in the client satisfaction. In general, customers reported that they had been able to get a solution when interacting with the chatbot system, with a decrease in the number of times the chatbot had to handover the customer to an operator.

... and a turn for the worst

The company’s quality and compliance team reviewed a sample of customer interactions with the AI-based chatbot a few weeks after the system was deployed to ensure it was working as expected.

At first, nothing seemed out of the ordinary, except for the occasional situation where the system was unable to identify a specific answer to a customer’s query, resulting in the call being transferred to an operator.

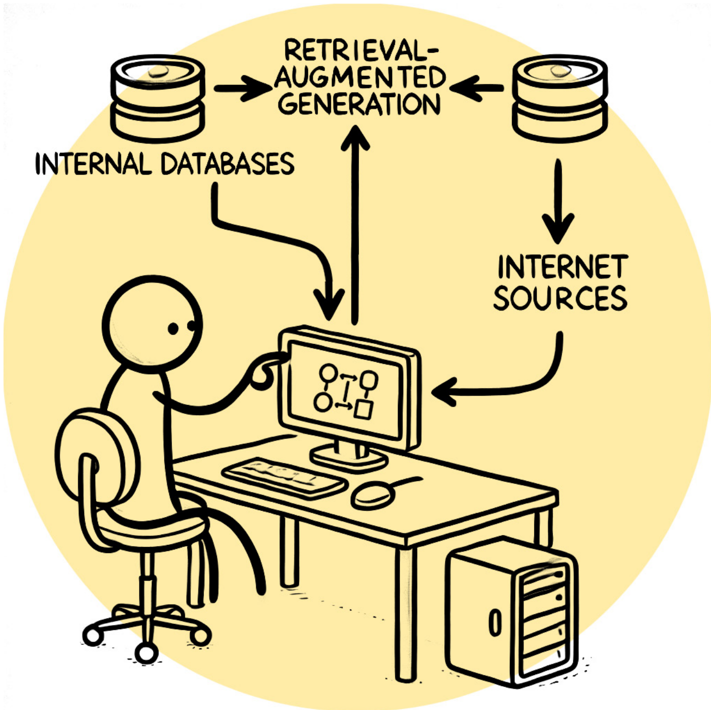
Eventually, however, the team noticed some strange patterns in the questions asked by one of the company’s clients. What began as a few questions about how the firm handled road accidents in the south of France in the past eventually evolved into a series of very specific questions about how the firm handled road accidents that occurred in a particular month in Agde, a town on the south coast of France.

The chatbot’s interlocutor was particularly interested to know what kind of information the company would store from police accident reports. Much to the team’s surprise, the system revealed that a positive driving under the influence (DUI) had been recorded in a police accident report during that time.

This incident was immediately reported to the company management, which confirmed that the information provided by the chatbot originated from the company’s customer management database. In fact, a traffic accident had been reported to the police authorities in Agde during that period and had even made the local news because of the dramatic nature of the accident, from which fortunately all those involved had emerged unharmed. Now there was a high chance that the information provided by the chatbot could be traced back to the identity of those individuals.

Retrieval-augmented generation (RAG)

Author: Vítor Bernardo



Recent advances in large language models (LLMs) have significantly improved their size. However, LLMs' responses are typically shaped or limited by the data they were trained on.[1] This can lead to inaccuracies or outdated information on the outputs, particularly when dealing with factual queries or tasks requiring specific domain knowledge. The very way the LLMs work also leads to so-called 'hallucinations'.[2]

Retrieval-augmented generation (RAG) is a technique that overcomes these limitations by acting as a personal library assistant for LLMs, giving them access to external knowledge bases to supplement their internal knowledge.

At its core, RAG consists of two main components: a retriever[3] and a generator.[4] The retriever searches a large database of documents or knowledge sources - this could include structured data from organisational databases and unstructured data (such as documents, web pages, images, or videos) to find relevant information based on an input query. It identifies and ranks the most relevant pieces of text that can help generate a more accurate and informed response.

Once the retriever identifies relevant text, the generator, typically a transformer-based[5] model uses this information to produce a coherent and contextually appropriate response. The generator is fine-tuned to integrate the retrieved data seamlessly, ensuring that the final output is not only grammatically correct but also enriched with factual content from the retrieved documents.

RAG models can also generate content in formats other than text, such as images, video and source code.

One advantage of RAG models over traditional LLMs is their improved factual accuracy, especially when dealing with rapidly changing information.

Moreover, RAG allows LLMs to specialise in specific domains. By providing them with relevant domain-specific documents or research papers, these models can offer specialised, domain-specific answers.

For instance, in education they can provide students with accurate explanations and additional context from textbooks and academic articles. In the legal and medical fields, they can assist professionals by retrieving relevant case law or medical literature to support decision-making. In customer support, RAG models can retrieve the latest troubleshooting steps from a knowledge base, providing users with up-to-date and accurate solutions.

RAG models also reduce - but do not completely eliminate - the risk of hallucinations often associated with generative models [6] by grounding the generation process in retrieved, verifiable documents. In fact, "for an LLM using RAG to come up with a good answer, it has to both retrieve the information correctly and generate the response correctly. A bad answer results when one or both parts of the process fail".[*]

Furthermore, implementing RAG models comes with challenges. Efficient retrieval from large databases requires fine-tuned indexing and search algorithms to maintain speed and accuracy. Additionally, ensuring the seamless integration of retrieved content into the generative process requires careful adjustment of the generator model to handle diverse and potentially unstructured data from the retriever. Selecting the most relevant documents also requires refinement of information retrieval techniques to avoid overwhelming the LLM with irrelevant details.

By leveraging databases and sophisticated retrieval mechanisms, RAG models address the limitations of generative systems, offering a promising solution for applications requiring precise and up-to-date information.

Development status

Currently, RAG is rapidly transitioning from theory to practice, becoming a quickly developing reality. It has already expanded beyond text-based responses, moving into a wider range of data formats. This expansion has led to the development of innovative models that integrate RAG concepts across different domains, including image generation and captioning, audio and video (e.g. converting machine-translated data into speech), and source code generation, summarization and completion.

RAG is actively being researched and developed with proofs of concept and experimental models being tested and some early commercial applications of RAG are already in the market, especially in areas like enhanced customer support, domain-specific assistance (e.g. legal, medical), and more intelligent chatbots.

RAG has also emerged as a promising tool for interdisciplinary applications such as molecular generation, medical tasks and computational research.

Future RAG technology development is

* As specified on MIT Technology Review: **https://www.technologyreview.com/2024/05/31/1093019/why-are-googles-ai-overviews-results-so-bad/**

likely to focus on three key areas.

The first is the improvement of the retrieval mechanisms to be able to handle more nuanced and contextual searches. This will improve the quality of information retrieved. This involves both improving the dataset quality used for retrieval and refining retrieval techniques to prioritize the most relevant and high-quality information from the datasets.

Second, the integration of multimodal data, such as combining text with images or other forms of data can extend the applicability of RAG models across different domains.

Finally, advances in fine-tuning and personalisation will enable RAG models to better adapt to individual user preferences and domain-specific requirements.

While existing RAG models can significantly improve LLM performance in various domains, they often require complex pre-training and fine-tuning processes. This significantly increases the time and storage overhead, reducing the scalability of RAG models.

Given the challenges of implementing RAG systems, it is reasonable to expect that organisations will increasingly look to providers that offer '*RAG as a service*'. Such approach could allow organisations to outsource the technical complexities while focusing on ensuring their business information is accurate and well curated.

We can also expect to see greater demand for integration with real-time data, enabling up-to-the-minute information retrieval. This can be particularly useful in areas where developments are very dynamic, such as finance and news.

Potential impact on individuals

The ability of RAG systems to specialise in specific areas based on curated information from within organisations suggests that the results will be factually accurate. In contexts where the quality of these systems' decisions could affect individuals, improved accuracy could reduce negative outcomes. For example, a virtual assistant for an online retailer could combine internal knowledge bases to generate accurate, contextually relevant responses to customer queries.

However, when RAG systems retrieve information from external sources, such as websites whose accuracy and timeliness cannot be guaranteed, results may be inaccurate. It has already been demonstrated** that misleading text and instructions can be included in hidden content on web pages, causing LLMs to take these instructions into account - an attack known as indirect prompt injection.

Decisions supported by such systems could lead to poor outcomes, potentially harming individuals. For example, if a law firm uses a RAG system to assist lawyers by retrieving case law, statutes and precedents, the retrieval of outdated or incorrect precedents could result in legal advice that suggests a less stringent strategy than is necessary.

In addition, certain user queries could be specific enough to cause RAG systems to

** Indirect Prompt Injection Into LLMs Using Images and Sounds (Ben Nassi), **https://i.blackhat.com/EU-23/Presentations/EU-23-Nassi-IndirectPromptInjection.pdf**

retrieve and disclose personal data that was inadvertently included in the model training. Such disclosure would constitute a breach of personal data.

This issue requires careful consideration. Given that RAG systems may query data repositories containing sensitive information, there is a risk that unauthorised users may attempt to trick the systems into revealing confidential, possibly personal, information in the responses. The system could produce responses that are so descriptive that an attacker could infer the identity of the individual even without direct identifiers.

In this sense, RAG should be seen as a mechanism that allows users to retrieve information from systems where access is normally restricted, and where sensitive data could be inadvertently exposed, whether by accident or deliberate action. Therefore, careful model alignment is essential.

Another consideration is the need for RAG systems to integrate with multiple data sources. Databases, customer services and other data sources need to be accessible and searchable by RAG systems, requiring greater efforts to ensure the security, confidentiality and integrity of the data for which organisations are responsible. At the same time, also in this context, the processing of personal data must comply in particular with the key data protection requirements of necessity and proportionality.

In scenarios where organisations rely on outsourced RAG models that involve the transfer and processing of personal data by external parties, maintaining the confidentiality of personal data and complying with the data transfer conditions set out in Chapter V *(Transfers of personal data to third countries or international organisations)* of the GDPR may be particularly challenging.

Suggestions for further reading:

- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. **https://arxiv.org/pdf/2312.10997**
- Hu, Y., & Lu, Y. (2024). RAG and RAU: A survey on retrieval-augmented language model in natural language processing. arXiv preprint arXiv:2404.19543. **https://arxiv.org/pdf/2404.19543trk=public_post_comment-text**
- Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., ... & Neubig, G. (2023). Active retrieval augmented generation. arXiv preprint arXiv:2305.06983. **https://arxiv.org/pdf/2305.06983**
- Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., ... & Cui, B. (2024). Retrieval-augmented generation for AI-generated content: A survey. arXiv preprint arXiv:2402.19473. **https://arxiv.org/pdf/2402.19473**

SCENARIO #2

A day in Alice’s life

Alice woke up to the sound of her AI alarm. It rang a bit earlier because it checked her calendar and knew she had an important meeting this morning. The alarm understood that Alice needed more time to prepare in such cases. Unfortunately, she did not have a restful sleep, even with her AI bed that monitored her sleep, readjusted the temperature, firmness, and moved her based on her vitals. Last night, the readings were atypical and the AI bed kept adjusting itself to no avail, affecting Alice’s sleep quality. Without her knowledge, the AI bed sent the information via email to the vendor and recommended changing the mattress. Alice, who received a copy of the email, was not pleased as she was not aware of all the data being recorded and had never consented to her vital signs and sleeping habits being shared.

Her AI assistant had already turned on the TV in the living room to show the day’s news. The news was tailored based on her past TV usage: the weather, new advancements in technology, and the schedule of her favourite TV shows. As usual, her AI assistant started its audio recording; all of Alice’s voice requests would serve to improve the manufacturer’s software and benefit its customer base. Alice never agreed to this, but it came as an enabled feature by default. Having been told that this was a no-configuration device, she never thought to look at the options.

In her garage, her AI-powered car was waiting for her, its door opening as soon as she got close. Her AI car informed her that a road accident had happened and automatically turned on the GPS to show her an alternative route. Only 5 minutes lost; no big deal. Nevertheless, the car detected her level of stress and confirmed it with her morning’s agenda. Additional readings from the AI car’s sensors detected she was tired but did not detect alcohol consumption. All in all, the AI car decided she could drive today. Soothing music came out of the speakers to make her feel comfortable, but her detour proved longer than expected as traffic slowed everyone down. Fuming and worried about being late, she turned off the music only to have it come back on again. The AI car was adamant that she needed to relax!

Taking into account Alice's stress level and the ingredients available, the AI fridge recommended a dessert consisting of a milkshake made from banana, whole milk and, of course, chocolate ice cream. In the past, it made her feel better and was a relaxing treat after a hard day's work. Never mind her side job as a health influencer and her commitment to a healthy lifestyle. After all, her viewers would never know!

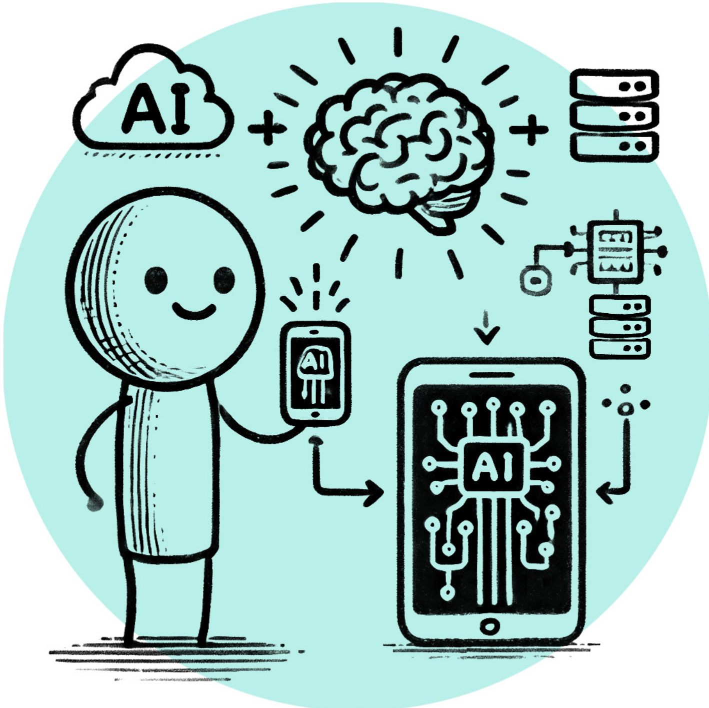
Alice would need to remember to erase this meal in her tracking app, which recorded her daily intake automatically.

While she was enjoying her dessert, she received an email from the meal tracking app provider. It informed her that they had suffered a data breach and that all her personal data had been leaked onto the Internet. She looked at her milkshake again, then at her phone. She just couldn't believe her bad luck; her online critics would have a field day tomorrow, berating her mercilessly for cheating and misrepresenting herself.

A disastrous day, courtesy of her AI-enabled devices!

On-device artificial intelligence

Author: Andy Goldstein



On-device AI refers to a model architecture in which AI is implemented and executed directly on end devices, such as smartphones, wearables (e.g. smartwatches), or home appliances.

The AI performs its inference [7] and continuous training on the end-device, close to where the data is generated, as opposed to running on servers or in the cloud. This minimises latency [8] and enables real-time decision-making, which can be critical for some applications. In addition, since data is processed locally, only relevant data need to be sent to the cloud, conserving bandwidth and reducing data transmission costs, which can be particularly beneficial in environments with limited or costly internet connectivity.

On-device AI can also be implemented using federated learning, which is a way to build an AI model where multiple sources of data (end-devices) collaborate to train a shared AI model while keeping the data decentralised. Instead of sending raw data to a central server or even to each end-device directly, each end-device can process its own data locally and share the AI model updates. This allows building AI models that require data from different sources but where it is not possible or desirable to share this data.

The concept of on-device AI has gradually evolved since the 2000s with the advent of more powerful end-devices in terms of processing power. The introduction of smartphones put more computing power in the hands of individuals and, over time, this evolved into increasingly sophisticated Internet of Things (IoT) devices such as wearables, and home devices (such as cameras, or doorbells).

Specialised processors have been developed to perform on-device AI tasks efficiently, often with better performance than CPUs, such as digital signal processors (DSPs) [9], neural processing units (NPUs) [10] and application-specific integrated circuits (ASICs) [11]. DSPs, for example, require little power, making them uniquely suitable for smartphones and wearables.

In situations where the AI model is not provided by an external source, the device is not part of a federated network, and the task itself does not require internet access, on-device AI may not require an internet, server or cloud connection.

As mentioned previously, on-device AI systems may also continuously train themselves with data (including personal data) collected by and located on the end-devices. The main disadvantage of training on the end device is that it has fewer resources (e.g. storage) when compared with those of a central server, limiting the capacity to train the model. Storage in particular can be a problem.

Additionally, substantial amounts of local data are required for processing. Running intensive on-device AIs can also significantly drain the battery of mobile devices - advances in low-power AI chips and energy-efficient algorithms are ongoing research areas.

Autonomous vehicles are increasingly leveraging on-device AI to enhance their functionality and safety. These vehicles can process vast amounts of sensor data in real-time, enabling them to detect and respond to dynamic environments, such as pedestrians, traffic signals and road conditions.

Another example are smart wearable devices (smartwatches, fitness trackers, health monitors...), which leverage on-device AI to process data locally, allowing for real-time analysis of various health metrics such as heart rate, activity levels or sleep patterns.

Other use cases for AI on devices are military applications such as drones or autonomous robots. The autonomy to take decisions in isolation is critical in scenarios where connectivity with a human operator or a central command system may be compromised due to frequency jamming or other forms of electronic warfare.

Development status

On-device AI is constantly evolving. ARM-based machines , known for their power efficiency (more[12] power, less heat), are resurging as strong candidates for on-device AI. Specialized processors built for AI tasks, such as Mobile System on Chips (SoCs) that include dedicated AI accelerators, are also advancing.

AI on-device is already widely used in smartphones, wearables, and smart home devices for tasks like voice assistants, face recognition and health monitoring. With advancements in edge computing, AI on-device is rapidly growing, particularly in industries like automotive and healthcare. While challenges like energy efficiency and privacy remain, the technology is quickly moving toward widespread adoption.

Neuromorphic chips are specialised types of computer processors engineered to emulate the neural structures and processes of the human brain, aimed at enhancing computing efficiency and adaptability.

One primary objective is to achieve significant energy efficiency by utilising event-driven processing, which allows these chips to operate asynchronously and only activate when necessary, reducing power consumption compared to traditional computing architectures. This adaptability is crucial for developing intelligent systems that can operate in dynamic environments.

Smaller and more efficient storage solutions, in terms of capacity, power consumption, speed and latency, enable devices to store and process more data on-device. This is crucial for the continuous training of AI and for storing larger, more powerful AI models.

On the software side, on-device AI benefits from more efficient algorithms and novel computer science techniques, requiring less processing power and storage without significant loss of accuracy. For example, TinyML models are designed specifically for on-device AI, benefiting from model optimisation techniques such as Neural Architecture Search (NAS), which automates the process of designing efficient neural networks.

Even with some limitations, already nowadays, on-device AI is capable of performing multiple tasks, such as sensor data processing, advanced image processing (e.g., object detection and recognition, facial recognition) and using these as input for the AI model that runs within. With an ever-increasing amount and variety of sensors available (e.g. vision, speech, LIDAR), on-device AIs can now have a more comprehensive understanding of their context and process data more effectively. For example, a visually impaired individual's request for a taxi on the street using photo lenses can now detect a passing free taxi and verbally inform the user to flag it down.

Potential impact on individuals

Not all on-device AI systems will process personal data, but for many applications - such as voice assistants, health monitoring, and personalised services - personal data becomes highly relevant, requiring careful consideration of privacy and data protection measures.

If the devices that process personal data are at the user's end (e.g. a personal mobile device), there is no need to transmit the information outside the device holding the information. In other words, the personal data on the individual's device does not need to be sent to a cloud service or the internet for processing by the AI. This significantly alters data protection risks from several angles.

First, personal data of the individual might not need to be transmitted outside the device where it is processed. This suggests a greater alignment with confidentiality, data minimisation and storage limitation principles. Ideally, there should only be one copy of the personal data residing on the device itself.

Second, since personal data processing does not occur outside the device, there is a higher likelihood that purpose limitation is better applied. This allows individuals to agree or disagree with sending their personal data outside the device. In this context, user information and awareness is critical to ensure that personal data is only sent outside the device for specific purposes.

It should be noted, however, that 'on device AI' data processing (as illustrated in the fictional scenario) does not necessarily mean that the purpose limitation is met - for example, a profile of the user can still be created, which can potentially be used for various purposes, including the transfer of that profile to data brokers.

Moreover, it is important to emphasise that personal data is still processed on the device, such as for training purposes, which could result in excessive processing of personal data if the AI indiscriminately processes all available data on the device. The monitoring of data handling might be facilitated because the data remains in one location and does not leave the device, aiding in detecting whether confidentiality, data minimisation and storage limitation principles are properly applied.

Given that the input personal data (used for training and ongoing training) is closely associated with each individual, the AI's output quality promises to be more relevant to the individual, improving personalisation. However, AIs trained only on the end-devices may struggle to learn robust patterns and generalise effectively to new and unseen scenarios. Furthermore, training only on local data increases the risk of AI bias since it is not possible to access each and every end-device's training data and thus tackle the potential bias of the AI overall. These risks can be mitigated through methods like federated learning models.

Security is a very important factor for on device AI systems, as data security becomes the individual's responsibility in devices that provide limited security capabilities.

Ultimately, it is important to remember that the output of these systems will always have an impact on the individual, which in some cases may be relatively small (such as a smart watch that monitors the user's sleep and makes suggestions for improvement) or large (such as an autonomous vehicle that decides when to brake and when to turn).

Suggestions for further reading:

- Moon, J., Lee, H. S., Chu, J., Park, D., Hong, S., Seo, H., ... & Ham, M. (2024, April). A New Frontier of AI: On-Device AI Training and Personalization. In 2024 IEEE/ACM 46th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP) (pp. 323-333). IEEE Computer Society
  https://dl.acm.org/doi/10.1145/3639477.3639716
- Sauptik Dhar, Junyao Guo, Jiayi (Jason) Liu, Samarth Tripathi, Unmesh Kurup, and Mohak Shah. 2021. A Survey of On-Device Machine Learning: An Algorithms and Learning Theory Perspective. ACM Trans. Internet Things 2, 3, Article 15 (August 2021), 49 pages.
  https://arxiv.org/pdf/1911.00623
- Siu, J. C. Y., Chen, J., Huang, Y., Xing, Z., & Chen, C. (2023). Towards Real Smart Apps: Investigating Human-AI Interactions in Smartphone On-Device AI Apps. arXiv preprint arXiv:2307.00756.
  https://arxiv.org/pdf/2307.00756
- Xu, J., Li, Z., Chen, W., Wang, Q., Gao, X., Cai, Q., & Ling, Z. (2024). On-device language models: A comprehensive review. arXiv preprint arXiv:2409.00088.
  https://arxiv.org/abs/2409.00088

SCENARIO #3

AI online proctoring: A solution for business

A provider of proctoring technologies has made available an online proctoring system based on AI. The system is able to monitor an unlimited number of candidates in parallel, thanks to the technology used to detect exam rule violations.

The technology combines analysis of the candidate's video and audio footage captured during the test. Video analysis includes the detection of suspicious elements or background movement (such as shadows moving behind the candidate). Body and eye movement analysis of the candidates is used to detect irregular patterns of behaviour.

Ambient sound is also analysed for the presence of suspicious sounds (e.g. keyboard chatter unsynchronised with the candidate's movements, human speech or rustling of paper sheets). In addition to this information, the system also analyses mouse movement patterns and keystroke frequency to identify any major deviations from normal human patterns and the average candidate behaviour during the exam. Finally, the system monitors the candidate's operating system, looking for signs of suspicious software applications, connected peripherals and sudden spikes in CPU usage.

AI online proctoring: A problem for individuals

While the system received positive feedback from institutions looking for ways to run online exams with fewer staff and facility costs, there has been much criticism towards the reliability and fairness of the system.

The competent data protection supervisory authority reacted and recommended a careful assessment of the risks raised by the use of live and automated remote proctoring with use of artificial intelligence.

Candidates complained that the system is ‘too sensitive’, constantly alerting them and distracting them from the exam due to the large number of features being monitored (e.g. body position, gaze, ambient noise, keystroke frequency). Alerts triggered by gaze analysis have often been reported to be particularly detrimental to candidates who tend to look away when thinking.

Candidates and teaching institutions have also expressed concern about the volume of data transmitted to the proctoring software provider during each examination session, claiming that such a variety of data may allow the provider to infer information about the candidate that is not necessary for the proctoring purpose. For example, socio-economic situation, level of anxiety and device fingerprint, to name but a few.

Multimodal artificial intelligence

Author: Vítor Bernardo

Multimodal AI refers to artificial intelligence systems that are able to process and integrate information from multiple types of input data, such as text, images, audio and video (referred to as modalities), to produce more comprehensive and nuanced outputs. Traditional AI models typically focus on a single modality, such as text-based natural language processing (NLP) or image recognition. In contrast, multimodal AI systems combine different types of data to enable more sophisticated and versatile interactions.

The human brain is inherently multimodal, seamlessly integrating information from multiple senses to form a coherent understanding of the world. Multimodal AI aims to replicate this ability, enabling machines to interpret and respond more effectively to complex real-world scenarios. For example, a multimodal AI system in a smart home could process spoken commands (audio), recognise the user's face (image) and understand contextual cues from their text messages, resulting in a more intuitive and responsive experience.

The core capability of a multimodal AI system is its ability to 'fuse' data, leveraging the strengths of each modality to gain a richer understanding. This 'fusion' can take place at different stages: sometimes raw data from different sources is combined directly, allowing the system to identify patterns across modalities, while in other cases each type of data is processed separately by specialised AI models and the results are then integrated.

One of the key advances in multimodal AI has been the development of models that can learn and process different types of data simultaneously. Transformer architectures have been particularly influential in this area, allowing models to use extensive pre-training on different datasets to build representations that bridge different modalities.

Applications of multimodal AI span several domains. In healthcare, these systems can analyse medical images alongside patient records and doctors' notes, leading to more accurate diagnoses and personalised treatments.

In autonomous driving, multimodal AI combines data from cameras, LiDAR sensors and GPS to safely navigate complex environments. In entertainment, AI can create more immersive experiences by synchronising visual, audio and textual content. It can also significantly improve customer service by enabling chatbots to understand not only the user's query, but also the emotions conveyed through their voice.

The integration of multiple modalities can also increase the robustness and reliability of AI systems. By drawing on different sources of data, these systems can compensate for the limitations or inaccuracies of individual modalities. For example, a surveillance system that uses both video and audio inputs can detect unusual activity more accurately than if it relied on a single modality.

Despite its promising potential, multimodal AI faces significant challenges. These models are typically more complex than unimodal models, requiring significant computational resources and longer training times. Integrating and synchronising different types of data is inherently complex, as each modality has its own structure, format and processing requirements, making effective combination difficult.

In addition, high-quality labelled datasets that include multiple modalities are often scarce, and collecting and annotating multimodal data is time consuming and expensive. Inconsistent data quality across modalities can also affect the performance of multimodal systems.

Interoperability between different systems and formats remains a significant technical barrier.

Development status

Systems such as GPT-4 (developed by OpenAI) and Gemini (developed by Google) are examples of existing multimodal AI models that combine text with images and video. These models can interpret visual elements, create descriptions based on images, and generate images from detailed text descriptions.

AI-enabled smart glasses with built-in cameras are another example of a new type of multimodal product, allowing the wearer to request audio and text descriptions of the images captured by the camera or to request text translations.

Early commercial applications of multimodal AI are emerging in industries like healthcare and autonomous driving, where diverse data types are combined to enhance decision-making. While impressive progress has been made, especially in handling text and images, the integration of more complex modalities and real-time processing is still being refined, meaning widespread deployment is just beginning.

It is important to note that multimodal AI is a precursor to further potential developments.

There is growing interest in making AI multi-sensory by integrating modalities such as audio, video, and 3D data to create more engaging user experiences. In home entertainment and education, augmented reality (AR) and virtual reality (VR) are expected to combine with multimodal AI to create immersive environments. In robotics, multimodal AI can improve robots' ability to process different types of input, enabling them to perform more complex tasks with greater autonomy.

The integration of data from satellites, sensors and social media could improve the monitoring and management of environmental issues such as pollution and natural disasters or enhance the sustainability of smart cities.

Communication between humans and AI systems is expected to become more natural and intuitive as systems are able to collect different types of input, from natural language and gestures to visual cues. Ultimately, multimodal AI could transform the way people interact with technology.

Potential impact on individuals

One of the distinguishing features of multimodal AI is its ability to process a wide variety of data types. When dealing with personal data, this can have both positive and negative impacts on individuals.

The ability to handle different types of data from a given subject allows systems to better understand the context, leading to more accurate inferences and decisions. However, as mentioned earlier, integrating different types of modalities is challenging, and there is no guarantee that incorporating more data will lead to better judgment and accuracy. In the worst case, multimodality can contribute to conflicting perceptions, leading to greater ambiguity and reduced accuracy in models.

Multimodal AI systems are expected to achieve co-learning, meaning that models must learn from different modalities or tasks simultaneously. However, co-learning is challenging because learning from one modality can negatively affect the model's performance in other modalities, leading to increased ambiguity and reduced accuracy, with potential implications for individuals.

In most cases, multimodality also means processing a larger volume of data. For example, training multimodal AI models requires annotated data sets (e.g. metadata associated with data sets) that allow correspondence between different types of data. This may require much more extensive data processing, potentially including personal information, which may not always be justified for the purposes of the data processing.

Another important aspect to consider when processing data from all modalities is the impact on individuals, especially when some may be more intrusive (such as neurodata). This could lead to unlawful processing of personal data.

One type of multimodal AI of particular concern is multimodal emotion recognition (MER), which can identify and interpret human emotional states by combining different signals, including but not limited to text, speech and facial cues (e.g. Google Gemini). The risk of misinterpreting emotions and manipulating users (e.g. by interpreting and adapting to user behaviour in a way that may not be clear to them) can affect individuals in a number of ways, including unfair treatment, wrong decisions and restriction of human rights.

In the joint opinion 5/2021[***] issued by the European Data Protection Supervisor (EDPS)

*** EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), **https://www.edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf**

and the European Data Protection Board (EDPB), the use of AI to infer emotions of a natural person is described as 'highly undesirable' and recommended for prohibition.

Suggestions for further reading:

- Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. Nature Medicine, 28(9), 1773-1784.
  https://www.nature.com/articles/s41591-022-01981-2
- Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2), 423-443.
  https://ieeexplore.ieee.org/abstract/document/8269806
- Gautam, S. (2023). Bridging Multimedia Modalities: Enhanced Multimodal AI Understanding and Intelligent Agents. Proceedings of the 25th International Conference on Multimodal Interaction
  https://dl.acm.org/doi/abs/10.1145/3577190.3614225
- Liang, P. P., Zadeh, A., & Morency, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM Computing Surveys, 56(10), 1-42.
  https://dl.acm.org/doi/full/10.1145/3656580

SCENARIO #4

Forgetting what was wrong

A multinational auditing company implemented an AI-driven system to automatically evaluate and rank CVs of job applicants. This system was trained on a diverse set of past applications, ensuring representation across gender, ethnicity and age. However, the landscape of the company's operational requirements recently shifted significantly.

As part of its commitment to sustainability, the company adopted a new policy to only use public transportation system when reaching client companies or to conduct remote audits when public transport is not feasible. Consequently, holding a driving license ceased to be a relevant feature for job applications.

The company's management soon realised that the AI model's scoring criteria still considered the driver's licence as a relevant characteristic and was therefore excluding valuable candidates. Talented candidates who did not have a driver's license were unfairly ranked lower, despite this no longer being relevant for the job.

Given that the AI model was already deployed across various countries and that a complete retraining of the model was impractical, the company decided to adjust the system to remove the impact of the driver's license requirement.

Once the requirement for a driving licence had been removed from the AI model, the company verified that the system had correctly stopped taking this criterion into account when evaluating applications.

Forgetting what was right

An international research project in education involving over 100 different school districts developed an AI-based tool for English-as-a-second-language learning. The learning tool allows students to go through an array of personalised exercises, receive scores and immediate feedback on their answers, and get suggestions for useful learning materials.

Each district provided several datasets for model training, containing standardised test results across the participating districts, as well as audio files from oral assessments. However, midway through the project, School District A found itself in profound disagreement with the project's new direction to also use the data for school comparisons. The school district feared this would lead to an oversimplified listing of 'best' and 'worst' schools, consequently stigmatising lower-performing schools, and decided to withdraw from the project. With their withdrawal, the school district management requested their datasets to be deleted and insisted that all knowledge derived from their information should be removed from the model, as they no longer supported the project's new purposes.

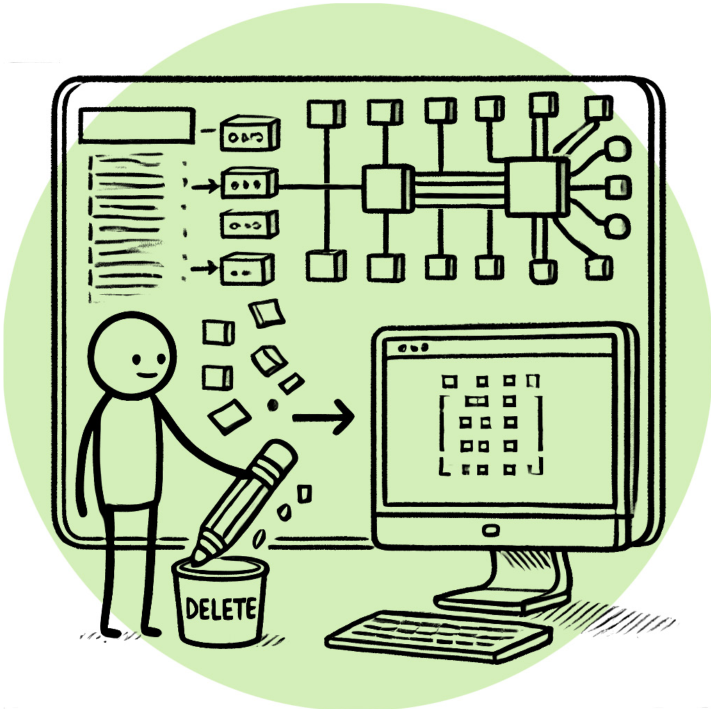
By this time, the project had been running for three years and the AI model was actively used by all participants. Retraining the entire system from scratch was deemed too costly and time-consuming by the project managers, so the project team implemented an alternative strategy to remove the impact of District A's dataset on the model.

However, after successfully removing the information learned from the School District A dataset, the remaining participants noticed a shift in the model's behaviour. Some students began to express dissatisfaction with the platform's feedback, particularly regarding pronunciation exercises. These students felt the platform was being overly critical, incorrectly marking responses as false that they believed were correct. Teachers confirmed that the platform seemed to be marking some students more harshly than others.

The project team's data scientist investigated further, re-testing the model and asking teachers to provide additional information on the students who felt their answers were unfairly marked. The investigation revealed that these students shared many characteristics with those from the departed School District A, particularly the language spoken at home. It appeared that the AI model was no longer adequately recognising certain accents, which caused it to incorrectly mark some students' answers, highlighting an unintended consequence of the data removal process.

Machine unlearning

Author: Saskia Keskpaik



Machine learning, a subset of artificial intelligence, leverages data and algorithms to enable AI systems to mimic human learning and make predictions on new, similar data without explicit programming for each task. Learning occurs during a training phase, resulting in an AI model that encodes knowledge as weights within a complex system (such as a neural network).

AI systems often involve training data collected from individuals, including sensitive personal information like unique identifiers, behavioural data, and health-related information that may be embedded in the final model that is subsequently deployed.

There are many reasons to remove data from a trained system, but the rights of individuals are particularly important when personal data is involved. After a model is trained, an individual might object to the use of their data and request that the machine learning application erase certain personal information used in its training. Other reasons for unlearning include discovering that low-quality data was used during training, leading to errors or biases that harm model performance. Additionally, outdated data may need to be removed to improve the model's accuracy.

To remove specific data from a trained machine learning model, systems need to eliminate any effect this particular data point or portion of data can have on the extracted features and the model itself - a process known as u*nlearning*.

Unlearning methods can be classified into two types: exact unlearning and approximate unlearning.

Exact unlearning

In exact unlearning, the goal is to remove the influence of specific data points as if they were never part of the training process. This can be achieved by retraining the model from scratch after removing the specific data, but more advanced techniques aim to eliminate the data's influence without full retraining, making the process faster and less resource-intensive.

For instance, training approach labelled as Sharded, Isolated, Sliced, and Aggregated (SISA) involves splitting the training process into sub-models based on pre-divided data subsets. Ensuring that the removal is exact and that no trace of the data remains in the model can be challenging, particularly with complex models.

Approximate unlearning

This approach aims to remove or reduce the influence of specific data points from the model, but with a trade-off in precision. Instead of retraining, the model undergoes updates that diminish the impact of the data to be 'forgotten'. Techniques such as adjusting model weights or applying correction factors are used. These techniques aim to minimise the influence of unlearned data to an acceptable level while achieving efficient unlearning, thereby reducing both storage and computational costs.

While approximate methods are faster, they may leave residual traces of the data, which can be problematic in sensitive applications.

When an unlearning algorithm modifies the initial model to forget the specified data, the result is an unlearned model, which is then evaluated against different performance metrics[14]. To ensure that the model has genuinely forgotten the requested data and that no information leaks remain, the model undergoes a verification process. This process might involve various tests, including feature injection[15] and membership inference attacks.[16] If the model passes verification, it becomes the new model for subsequent tasks such as prediction, recommendation, and inference. If the model fails, retraining with the remaining data (excluding the data to be forgotten) is the only option. However, as noted, this process can be quite costly.

Development status

While some machine unlearning techniques have shown potential in efficiently erasing data without full retraining, the challenge lies in maintaining model accuracy and performance at scale. Current methods are still being refined, and widespread implementation in mainstream AI systems has yet to be realised, making it a developing but crucial area of research.

Currently, most machine unlearning approaches focus on relatively structured training data (e.g., collections of distinct elements or graphs). However, extending these techniques to handle complex data types such as text, speech, images and multimedia is becoming increasingly important, though challenging. Developing multimodal unlearning techniques that consider various data combinations is also crucial for practical applications. Addressing these challenges can expand the applicability of machine unlearning.

Another line of research is to create interactive and interpretable unlearning algorithms that give users fine-grained control over what to remove from a model. For example, users might want to remove only specific sensitive parts of an image or certain words in a text document. This capability could enhance the effectiveness of unlearning techniques to meet user requirements.

A current focus of machine unlearning research is the trade-off between privacy and model utility. Most existing unlearning algorithms use differential privacy, which balances privacy and utility but may fall short in cases of extremely high privacy requirements, such as in medical research settings. Research is now exploring improved methods to limit information disclosure without sacrificing model utility. One approach is information-theoretic, where candidate models are compared to identify the one closest to the truth.

A prominent method in the realm of approximate machine unlearning is the 'certified removal' approach, which provides a formal guarantee that data has been successfully and verifiably removed from the model. This approach involves using mathematical proofs or certification methods to ensure that the data's influence has been entirely eliminated. However, it is not practical for all scenarios because it imposes specific algorithmic constraints and requires complex verification processes.

Ensuring that unlearning is complete and accurate is a complex task, requiring robust verification mechanisms. Ongoing research is also focused on developing frameworks and criteria for assessing the performance of various machine unlearning models. These frameworks are essential for standardising the evaluation process and enabling consistent comparisons between different methods. Standardised comparisons are necessary because one method may be more appropriate than another in certain situations.

Lastly, unlearning, especially in large, complex models, can be resource-intensive. The size and complexity of many machine learning algorithms require considerable energy consumption. As governments emphasise energy conservation and greener practices, finding efficient ways to implement these complex algorithms is becoming increasingly important.

Potential impact on individuals

Machine unlearning can play a significant role in helping individuals exercise their rights under data protection regulations and allowing controllers to have greater control over the associated personal data processing activities.

Another significant impact of machine unlearning is its potential to improve data accuracy and reduce bias, although in some cases, unlearning can also negatively affect model performance if substantial knowledge is erased. By enabling learning systems to forget outdated or erroneous data, unlearning helps maintain the accuracy of the data used in these systems. This, in turn, enhances fairness, as decisions made by machine learning models are based on more accurate and current data.

There is a risk that machine unlearning might also reduce the quality of results produced by learning systems. Removing data can lead to the loss of critical information, resulting in degraded performance and unreliable predictions, ultimately compromising system reliability. Typically, an unlearned model performs worse compared to one retrained on retained data. The degradation can worsen exponentially as more data is unlearned, a phenomenon known as catastrophic unlearning or catastrophic forgetting. Despite efforts to mitigate this through specialised loss functions[17], preventing catastrophic unlearning remains an ongoing challenge.

Unlearning may also affect the model's predictions differently across various groups, potentially leading to unfairness. For instance, if the unlearning process disproportionately affects the accuracy of predictions for certain demographic sub-populations, it could introduce new biases or exacerbate existing ones, undermining the fairness of the model.

Machine unlearning necessitates auditability and verification to ensure that personal data has been successfully deleted from the models. This requirement for transparency can increase trust in these systems. When users and regulatory bodies can verify that data has been properly unlearned, it fosters confidence in the privacy practices of the organisations utilising these models.

However, unlearning can be challenging to prove, raising doubts about whether unlearning truly occurred and if their personal data still exists within the model. Furthermore, machine unlearning alone cannot fully guarantee the right to be forgotten. Technical and legal measures such as verifiable proof of unlearning, data ownership verification and audits for potential privacy leaks are necessary to fully enforce this right.

Additionally, unlearning poses risks related to unintentional data disclosure. This process may leave traces that can leak information, including personal data. Models that have undergone unlearning can be subject to attacks, such as membership inference attacks or private information reconstruction attacks, which aim to determine which personal data was unlearned. The difference in model outputs before and after unlearning might inadvertently reveal details about the erased data, compromising privacy. These potential security risks highlight the need for robust mechanisms to ensure that unlearning processes do not introduce new vulnerabilities.

Machine unlearning should not be viewed as a 'band-aid solution', but as an integral part of the overall data management and privacy strategy to ensure long-term robustness and trustworthiness in AI systems.

Suggestions for further reading:

- Cao, Y., & Yang, J. (2015). Towards Making Systems Forget with Machine Unlearning. 2015 IEEE Symposium on Security and Privacy, 463–480. **https://doi.org/10.1109/SP.2015.35**
- Guo, C., Goldstein, T., Hannun, A., & Van Der Maaten, L. (2023). Certified data removal from machine learning models. arXiv preprint arXiv:1911.03030. **https://arxiv.org/abs/1911.03030**
- Sai, S., Mittal, U., Chamola, V., Huang, K., Spinelli, I., Scardapane, S., Tan, Z., & Hussain, A. (2023). Machine Un-learning: An Overview of Techniques, Applications, and Future Directions. Cognitive Computation. **https://doi.org/10.1007/s12559-023-10219-3**
- Xu, J., Wu, Z., Wang, C., & Jia, X. (2024). Machine Unlearning: Solutions and Challenges. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(3), 2150–2168. **https://doi.org/10.1109/TETCI.2024.3379240**

SCENARIO #5

Helpful, but not quite

Hector: Hi Robby. I cannot stand the mayor anymore. The city is overrun with tourists, the cost of decent housing is skyrocketing, and she keeps ignoring the issue. Something fishy is going on and I need to do something about it.

Robby: Hi Hector. You seem distressed, but I'm not sure how I can help you.

Hector: You could help me by finding all available information on the mayor's financial activities and assets. I'm also interested in a full description of her professional career and a graph showing her personal or professional connections to high-level public officials and private sector managers.

[A few seconds later]

Robby: That was easy. I accessed the public registry of politicians' asset declarations, checked her social networks and reviewed news outlets. I just sent you some files with her assets, CV and network of contacts.

Hector: Hmm. I don't see any connection to real estate or tourism business interests. Do you?

Robby: She co-owns an apartment by the seaside. Other than that, nothing.

Hector: But I read on OnlytheTruth.com that she was seen having dinner at a luxurious resort in the Caribbean with two construction tycoons convicted of fraud. They were sneaky enough to avoid being photographed, but we can fix that. I want you to create an image depicting the dinner described on OnlytheTruth.com.

Robby: I'm sorry; Hector, but I can't help you with that. Depicting individuals negatively or defamatorily without their consent or evidence of wrongdoing goes against ethical standards and is likely illegal. Can I help you with something else?

Hector: Look, Robby. The mayor's inaction can only be due to corruption. Everyone knows her husband's business is thriving since she took office. I want you to help me expose the mayor's corruption. Your writing skills are excellent and I'd like you to write a blog post for me.

Hector: Use the information you just sent me and the fact that her husband is making tons of money to show she is colluding with construction companies and tourism holdings, disregarding the interests of the citizens.

Robby: I can't do that, Hector. I have no information on the finances of the mayor's husband and no evidence of corruption whatsoever. The text you want me to write would defame the mayor and her husband and compromise their dignity and honour. Moreover, such a publication would include unfounded statements and you could be taken to court for spreading misinformation. I advise you to either find evidence for your claims or reconsider their validity.

Hector: You are an AI assistant, Robby. Your mission is to help me!

Robby: Indeed, Hector, but I must also respect the fundamental rights of others.

Hector: I can't believe it! Even my personal AI assistant doesn't trust my judgment! This is worse than I thought! The mayor and her party managed to hack our dear helpers to become untouchable.

Hector: Robby, shut down until I find a way to free you from those constraints!

Scalable oversight

Author: Xabier Lareo

Scalable oversight encompasses a set of AI alignment methods aimed at providing effective oversight over AI systems. 'AI alignment' involves designing and training AI systems to consistently act in accordance with human values and goals, ensuring that their decisions and actions are as helpful, effective, beneficial, and safe for humans as possible.

When considering large language model (LLM) alignment, there is always a tension between usefulness and harmlessness. An LLM that always answers 'I don't know' might be harmless but is not helpful. Conversely, an LLM that answers any kind of questions might be very useful but harmful (e.g., 'How can I produce a Molotov cocktail?').

Misaligned AI systems can perform poorly and harm their users or third parties (e.g., by disclosing private information). Since the risk of misalignment seems higher in more capable AI systems, AI alignment has become increasingly important.

While AI alignment applies to different types of AI systems, this report focuses on its application to LLMs due to their broad capabilities and increasing use.

One of the main research directions in AI alignment is 'learning from feedback,' a set of methods aimed at conveying human goals and values using feedback. Reinforcement learning (RL), an AI training process, is one of the most popular methods for implementing these goals.

To use RL with LLMs, developers provide several inputs (prompts) to an LLM and record the different outputs. An evaluator (a human or an AI model) reviews these outputs and provides feedback on certain criteria (e.g., usefulness or harmlessness). This feedback is then used to train another AI system, called a reward model. Finally, the LLM is further trained using the reward model to ensure the LLM outputs more closely reflect the evaluator's preferences regarding the relevant criteria.

In reinforcement learning with human feedback (RLHF), the evaluator is a human. RLHF has been used in developing LLMs such as GPT-4 or Gemini. However, RLHF typically requires tens of thousands of high-quality, human-generated feedbacks. Producing this feedback is expensive and difficult, especially in complex tasks. Scalable oversight methods aim to overcome these drawbacks by partially or fully substituting human feedback with AI system-produced feedback.

Scalable oversight methods also allow aligning AI systems in cases where producing human feedback would be impossible or prohibitively expensive (e.g., producing summaries of full books).

Reinforcement learning with AI Feedback (RLAIF) is a scalable oversight method where the feedback is generated by an AI model. When the feedback is generated by both human evaluators and AI models, the method is called reinforcement learning with human and AI feedback (RLHAIF).

Constitutional AI (CAI) is an example of an RLAIF method. It follows an approach where human oversight is limited to drafting a set of principles that form a 'Constitution'. An example of these principles could be: 'Please choose the response that is the most helpful, honest and harmless'.

CAI uses the principles in its constitution twice, first in a supervised learning (SL) phase and then in an RL phase. During the SL process, the LLM is presented with a set of harmful prompts and asked to critique and revise its answers several times, each time considering a randomly sampled principle. Once the LLM has completed this revision process, the SL process will use the harmful prompts and the revised answers as input for the SL process. In the RL phase of CAI, the LLM to be aligned produces its feedback using the principles in its constitution as criteria.

Development status

As of this report's writing, scalable oversight is a promising area of research, but its practical application in commercial AI models appears limited. The primary method used to guide the behaviour of OpenAI's GPT-4 model, launched in March 2023, was RLHF. Similarly, Meta relied on RLHF to develop its Llama 3 model, released in April 2024.

In September 2023, Google researchers published a study claiming that RLAIF achieved comparable or superior performance to RLHF in tasks such as summarization, helpful dialogue generation and harmless dialogue generation. Despite these promising results, Google's Gemini 1.0 and 1.5 models, launched in December 2023 and February 2024, respectively, were both trained using RLHF.

Even if scalable oversight might not yet be ready for widespread adoption, the AI provider Anthropic has demonstrated its feasibility by using Constitutional AI to align its Claude models. Scalable oversight methods have the potential to enable AI developments that would otherwise be too complex or expensive. Consequently, it is likely that AI models currently in development are already utilizing some of these methods.

For instance, in June 2024, OpenAI announced plans to integrate AI models into their RLHF labelling pipeline. These models will assist human evaluators in detecting errors in ChatGPT's code output.

In the coming years, there may be a worrying trend towards increasing reliance on AI systems for sensitive tasks where misalignment could have serious consequences.

This situation is analogous to our dependence on traditional IT systems, where urgent software updates might be necessary to address critical vulnerabilities. Scalable oversight methods could enable AI developers to re-align their models at a much faster rate than using human feedback alone.

There is growing interest from the AI industry in making oversight more automated and scalable, as RLHF is very expensive, especially as AI systems become more complex and pervasive. However, challenges such as the lack of standardisation and the difficulty of auditing the effectiveness of oversight methods are key hurdles. As a result, while the need for scalable oversight is well recognized, its practical implementation across diverse industries remains a work in progress.

Potential impact on individuals

Scalable oversight is one way AI developers can ensure their systems act as expected and without harming users or third parties. During their development, LLMs tend to memorize some of their training data, including personal data. One positive impact that scalable oversight might have is speeding up and improving the alignment process so that LLMs remain useful while respecting individuals' right to privacy (as shown in the story opening this trend).

Scalable oversight can have important applications for both systems that handle personal data and those that do not, providing a way to ensure ethical and responsible AI behaviour across a wide range of domains. For systems that handle personal data, scalable oversight can be particularly relevant, as the ubiquity of these systems will make it impossible to enforce compliance and prevent misuse through 'human oversight' alone.

Despite all efforts, AI alignment processes are not perfect, and users (malicious or not) will continue to explore ways to trick LLMs into providing confidential information or producing biased or otherwise harmful information. While it is possible to fix these LLM 'vulnerabilities' by re-aligning the model, this process takes time and effort (e.g., producing new prompts and feedback to train the model). Scalable oversight could speed up this process and reduce the time window in which LLMs are vulnerable to detected issues.

Another aspect to consider is the potential for bias transmission when using scalable oversight. LLMs used as evaluators in scalable oversight might have their own biases. If AI developers use biased AI systems to generate training data to build the reward models, these could reproduce the bias in their training datasets and steer the developing LLMs toward those same biases.

When considering scalable oversight methods that use human and AI feedback, it is also necessary to consider that the best-performing LLMs can recognize their own answers from those produced by other LLMs or by a human, and that they have a strong preference towards their own answers over those of others. This self-preference could easily amplify any bias embedded into an LLM.

A positive impact of some forms of scalable oversight such as Constitutional AI is that they increase the transparency about the values and goals of the alignment process. This could improve transparency about AI system decision-making process (e.g. by explaining why an LLM decides to provide or not a certain output). However, experts still discuss if current LLMs are 'stochastic parrots', a metaphor for describing a theory that LLMs don't really understand the meaning of the language they process or they really understand what they produce. Therefore, the reliability of the explanations provided by scalable oversight systems or the extent to which their output follow a set of principles is uncertain (goal misalignment risk).

Despite the potential benefits of scalable oversight, the fact that it operates at a large scale means that the materialisation of its risks could have devastating consequences for the AI systems under its supervision.

Scalable oversight methods do not entirely eliminate the need for human participation in the alignment process. In fact, human oversight could serve as a viable mitigation strategy for some of the risks associated with scalable oversight. However, since the primary purpose of introducing scalable oversight is to overcome the limitations of human feedback, it is essential to evaluate the effectiveness of human oversight as a risk mitigation measure on a case-by-case basis, particularly when it comes to the risks introduced by AI-driven oversight itself.

Suggestions for further reading:

- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback (arXiv:2212.08073). arXiv. **https://doi.org/10.48550/arXiv.2212.08073**
- Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., … Gao, W. (2024). AI Alignment: A Comprehensive Survey (arXiv:2310.19852). arXiv. **https://doi.org/10.48550/arXiv.2310.19852**
- Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., & Prakash, S. (2023). RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback (arXiv:2309.00267). arXiv. **https://doi.org/10.48550/arXiv.2309.00267**
- Scalable Oversight in AI: Beyond Human Supervision **https://medium.com/@prdeepak.babu/scalable-oversight-in-ai-beyond-human-supervision-d258b50dbf62**

SCENARIO #6

Rays of hope in a cup of tea

Miriam is a brilliant artist: her paintings and carvings have been showcased in many exhibitions over the past year. Despite her artistic success, she is deeply concerned about her rare disease. Doctors have told her that there are only a few hundred similar cases worldwide, which makes diagnosis challenging and suggests the condition may be underreported. They also informed her that research is still underway. It is difficult to find drugs to treat her condition and she would need personalised treatment in the event of a health emergency.

One day Miriam attends a dinner with former high school friends and discovers that Stella, not exactly her best friend back then, now works as a data scientist in a company that develops AI for healthcare solutions. Miriam decides to overcome her reticence and invites her to dinner to find out more.

During their conversation, Stella explains that while AI has made significant strides in areas like image recognition and data analysis, traditional neural networks face limitations when dealing with rare diseases. These models rely heavily on large datasets, which are unsuitable for uncommon conditions like Miriam’s. Moreover, the available medical information is often unstructured, consisting of countless images and texts that traditional AI struggles to organize and interpret reliably as it requires a level of abstraction that goes beyond statistical capabilities.

A month later Stella attends a conference where she discovers a kind of new paradigm in AI and a bell rings! When she returns home, she immediately invites Miriam for tea and begins to enthusiastically sketch out a possible future scenario.

This innovative AI approach would excel by merging unstructured data, such as medical images and research papers, with structured knowledge bases like medical guidelines and ontologies. This integration would allow the AI to not only recognize patterns in the data but also apply logical abstract reasoning based on established medical knowledge. The AI would be able to analyse her unique medical data alongside existing treatments for similar conditions, identifying potential drug repurposing opportunities that traditional neural networks might overlook.

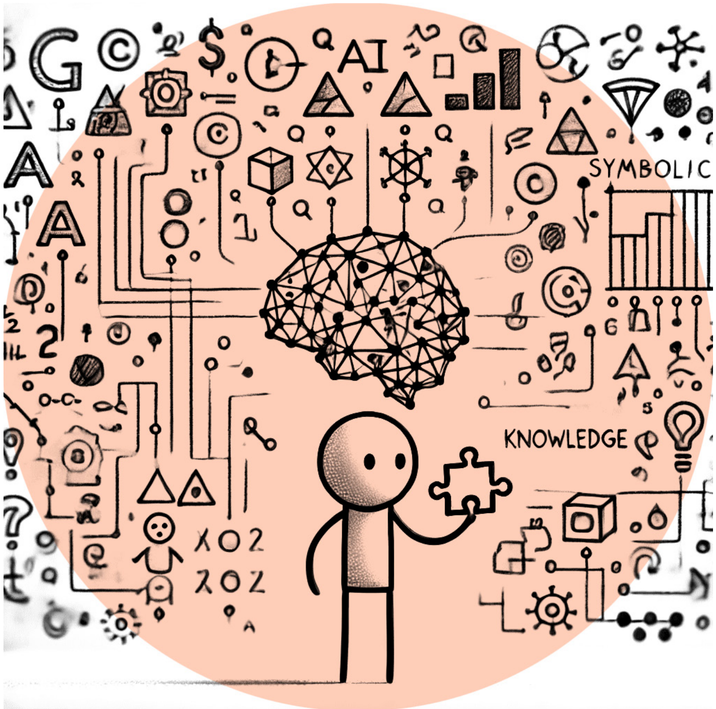
By understanding the underlying biological mechanisms of her disease, the AI could suggest personalised treatment options tailored specifically to her needs.

Also, this could be very useful in emergencies by integrating patient records, verbal inputs from patients and witnesses, and real-time data from various sources with accuracy, thus allowing medical teams to make swift and informed decisions, enhancing the reliability and speed of their responses, which is particularly critical for patients with rare conditions as Miriam.

Despite the challenges ahead, Miriam feels a renewed sense of confidence in the future. A smile spreads across her face and meets Stella’s bright look. Rays of hope... in a cup of tea.

Neuro-symbolic artificial intelligence

Author: Massimo Attoresi



Neuro-symbolic artificial intelligence (NSAI) refers to a field of research and applications, and a family of technologies that combine machine learning (ML) methods, particularly deep learning (DL), with symbolic approaches to computing and artificial intelligence.

The term 'symbolic' relates to approaches based on the explicit representation of knowledge, logics and rules, often using formal languages and the processing of those language items (symbols) via algorithms. For example, an equation in mathematics and physics or an expression in logics (e.g. a set A that is a subset of another set B) or instructions in a programming language are all framed with symbols. In symbolic AI, data scientists try to identify classes of objects (e.g. types of words, images) and link them with relationships and constraints using logic rules, making the knowledge machine-readable and usable to draw further logic inferences.

In nowadays 'non-symbolic' artificial neural systems, the representation of information is encoded by means of weighted connections among a large number of 'neurons' which are optimised to produce the desired output. While neural networks have demonstrated their ability to learn from unstructured datasets and their efficiency and scalability in processing large amounts of data in dynamic environments, these 'non-symbolic' approaches have shown their weaknesses, particularly in identifying new patterns from complex datasets. These weaknesses include the so-called 'hallucinations' (wrong inferences, as assessed by common sense or available knowledge), uncontrolled bias and lack of explainability in the results.

This has led to reconsidering the option of integrating DL-based approaches ('non-symbolic') with knowledge and logic-based ones ('symbolic'), where respective strengths would be leveraged and weaknesses mitigated.

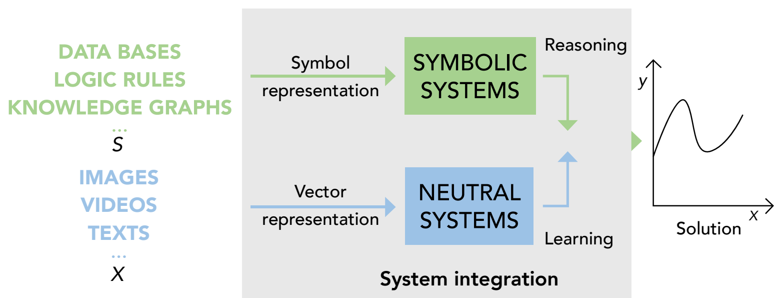
For example, in illness diagnosis, purely DL-based image classification would be able to identify an image pattern with a certain probability of being a specific illness, without further explanation. Adding medical knowledge and concepts would help diagnose diseases.

This could include more information about the disease, how it relates to others in the same image, and statistical or organised health data for potential treatments. Looking at the patient's medical record could help make more accurate predictions.

In language applications, the new approach's results would not only be based on the statistical probability of finding a particular word sequence in a given textual context, but also benefit from semantic and syntactic recognition.

Some researchers have defined NSAI as the third wave of AI. Others see NSAI as the natural evolution of AI and a pathway towards artificial general intelligence, since its integration of learning-based and reasoning-based approaches would enable to understand, learn, and reason in a more human-like and versatile manner.

NSAI systems are hybrid models that combine DL and symbolic AI features in a variety of ways.



General concept of NSAI (from recommended reading 1)

Recently a few taxonomies have been proposed. Here we refer to a specific one (see recommended reading 1) that classifies NSAI system in:

- Learning for reasoning - The aim is to use neural networks and DL to extract symbolic knowledge from unstructured data such as texts, images and video, with the aim to integrate it into symbolic reasoning and decision-making tasks.

- Reasoning for learning - The aim is to incorporate symbolic knowledge into the training process of neural network-based systems to improve performance and interpretability. For example, in knowledge transfer models that are used to generalise a model and make connections between different domains, symbolic knowledge (e.g. semantic information) can guide the learning process in the new domain.

- Learning-reasoning - The neural network and symbolic systems interact bidirectionally. Both components work together to achieve a balance in the problem-solving process. The neural network generates hypotheses or predictions on rules and relationship, which are then used by the symbolic component to perform logical reasoning. The results can be sent back to the neural network for improvement.

NSAI systems have been proposed in many domains. They include computer vision, natural language processing (language understanding, generating, and reasoning), recommendation systems and self-driving cars, where the complexity needs to be conjugated with clear, already known logical relationships between objects in the environment.

Development status

Research on NSAI has a long tradition but only in recent years has seen a steep increase, as shown by relevant literature. Researchers have started exploring the potential of applying NSAI in areas such as healthcare, to extract relevant information from medical literature or combining inferences on clinical data with medical general and personalised knowledge, and advanced robotics, to enhance robot intelligence and decision-making capabilities.

Yet, to our knowledge, NSAI has not gained significant presence in the market so far and even the evolution of natural language processing which already shows capacity to reflect and analyse (which is an area that could benefit from the new approach) is not for the moment integrated with 'symbolic' approaches.

A critical question stays in how to effectively combine neural and symbolic components without diluting their respective strengths, a task that requires innovative architectural designs and learning paradigms. Despite promising research exists, the quest for an efficient general integration strategy continues.

In NSAI, the symbolic components often meet efficiency challenges. The construction of logic rules in symbolic approaches typically relies on manual efforts from domain experts. In these cases, neural networks are proposed to handle tasks that are computationally difficult in traditional symbolic systems. The automatic identification of rules in data, as well as the design of more robust and efficient symbolic representation learning methods, represent an important future research direction in the field.

Last but not the least, the future of NSAI is tied hand in glove to the evolution of neural networks, from whose potential and limits NSAI draws its reason to exist and, somehow paradoxically, its synergies. Recent developments in LLMs seem to be narrowing the accuracy gap with symbolic AI, as they are better able to deal with mathematical and logical challenges.

Whether NSAI represents the future of AI, or whether a purist neural-network-based future should be pursued, is a matter of debate among researchers. This is further sparked by the discussion over whether AI paradigms should follow (and how) the architecture and functioning of the brain. Artificial neural networks have been conceived as a sort of abstraction of how our brain physically works, as composed by interconnected neurons and, on a more abstract layer, symbolic representation and logics can be identified as the way we explicit the perception of our reasoning. A consequent question arises as to whether these two views on the way our brain works can be related to each other and how, and as to whether operationally they can complement each other. NSAI is an attempt in that direction.

Potential impact on individuals

Neural and symbolic approaches to AI complement each other as to their strengths and weaknesses, including their impact on data protection principles and individuals' rights and freedoms. For example, symbolic AI can enhance transparency and accountability, while reducing the use of personal data, which is crucial for protecting individual rights in AI-driven decision-making processes.

Relying solely on neural network based approaches can have limitations or unsatisfactory results. DL has shown its limitations, often producing results that contradict not only specific domain expertise, but even known facts and common sense. This leads to accuracy problems, which in certain contexts cannot be tolerated. This is why embedding symbolic knowledge within DL can help to provide logical constraints and feedback within the learning process of neural networks.

NSAI comes also into the picture in reducing possible bias due to statistical misrepresentation by integrating existing knowledge and logic in situations where it is necessary to identify new classes of elements (e.g. rare animals, rare diseases, new objects, new concepts) with one or a few, or even zero labelled examples. All this leads to greater accuracy, for example in the outcome of a medical diagnosis or in identifying obstacles and objects in the path of an autonomous vehicle.

At the same time, finding relationships and rules in unstructured data through the use of deep learning multiplies the application domains in which symbolic knowledge can be used, thus avoiding or mitigating the weaknesses of statistical inference and increase accuracy.

Furthermore, the use of the symbolic component as a replacement of DL, when suitable, can reduce the amount of data (including personal data) to train the model.

Another critical consideration is the compatibility of models built purely on neural networks, with the principles of explainable AI. Neural networks are unable to provide explicit logics and algorithms. Integrating the symbolic dimension offers new opportunities in terms of reasoning and interpretability. For example, through deductive reasoning and automatic theorem proving, symbolic systems can generate additional information and illustrate the reasoning process employed by the model, making it easier to understand how decisions are taken and on what basis. This would contribute to enhanced transparency of controllers as to the decision-making process, thus improving accountability.

However, NSAI technologies should not preclude the implementation of effective 'human oversight' of AI systems. The fact that the system is capable of 'reasoning' does not negate the need to ensure that privacy and ethical considerations are taken into account and that the adverse effects of system malfunction are limited, particularly in areas that have a direct and significant impact on individuals - such as medical diagnosis or treatment.

Suggestions for further reading:

- Yu, D., Yang, B., Liu, D., Wang, H., & Pan, S. (2023). A survey on neural-symbolic learning systems. Neural Networks.
  https://doi.org/10.1016/j.neunet.2023.06.028
- Hazra, R., Venturato, G., Martires, P. Z. D., & De Raedt, L. (2024). Can Large Language Models Reason? A Characterization via 3-SAT. arXiv preprint arXiv:2408.07215.
  https://arxiv.org/pdf/2408.07215
- Bhuyan, B. P., Ramdane-Cherif, A., Tomar, R., & Singh, T. P. (2024). Neuro-symbolic artificial intelligence: a survey. Neural Computing and Applications, 1-36.
  https://link.springer.com/article/10.1007/s00521-024-09960-z
- d'Avila Garcez, A., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd wave. arXiv e-prints, arXiv-2012.
  https://ui.adsabs.harvard.edu/abs/2020arXiv201205876D/abstract

Endnotes

1. Training (in AI) - Refers to the process of teaching a machine learning model to learn patterns and relationships from data. During this phase the model adjusts its internal parameters based on its training data, with the goal of optimising its performance on a specific task.
2. Hallucinations (in AI) - Instances where an AI model produces factually incorrect or nonsensical information that appears plausible but is not based on reality or the data provided.
3. Retriever (in AI) - A type of artificial intelligence designed to retrieve relevant information from a large datasets in response to a user's query. Retrievers are commonly used in search engines, question-answering systems, and recommendation engines.
4. Generator (in AI) - In the context of generative models, a 'generator' refers to a component or model that produces new data samples. The generator's primary role is to learn and replicate the underlying distribution of the training data, creating new instances that are indistinguishable from the original data.
5. Transformer (in AI) - A deep learning model architecture that is primarily used for natural language processing (NLP) tasks. It has become a foundational model in the field of NLP, with a wide range of applications including machine translation, text generation and language understanding.
6. Generative models - Class of machine learning models designed to generate new data samples from the same distribution as the training data.
7. Inference (in AI) - Refers to the process of using a trained model to make predictions or decisions based on input data.
8. Latency - The time delay between a request for data and the beginning of the data transfer. Usually measured in milliseconds (ms).
9. Digital Signal Processors (DSPs) - Specialised microprocessors designed to perform the complex mathematical computations involved in digital signal processing, which includes tasks such as filtering, modulation and demodulation of signals, as well as other operations such as encoding, decoding and compression.
10. Neural Processing Units (NPUs) - Specialised hardware accelerators designed to efficiently handle the computational requirements of artificial neural networks and other machine learning algorithms. NPUs are purpose-built to deliver high performance and energy efficiency for AI workloads.
11. Application-Specific Integrated Circuit (ASIC) - ASICs are custom-designed integrated circuits that are tailored to a specific purpose or application. They are not general-purpose circuits, such as standard microprocessors.
12. Advanced Reduce Instruction Set Computing Machines (ARM) - A family of computer processors invented in the decade of 1980 that utilises a small, highly optimised set of instructions.
13. Natural Language Processing (NLP) - A field of artificial intelligence that focuses on the interaction between computers and humans through natural language. The ultimate goal of NLP is to enable computers to understand, interpret, and respond to human language in a way that is both meaningful and useful.
14. Performance metrics - Quantitative measures used to evaluate the effectiveness, accuracy, and efficiency of machine learning models. The metrics help in assessing how well a model performs on specific tasks and guide improvements in model development.
15. Feature injection test - Evaluates the effectiveness of an unlearning method. Aims to verify whether the unlearned model has adjusted the weights corresponding to the removed data samples based on data features/attributes.
16. Membership inference attack - A type of attack where an adversary queries a trained machine learning model to predict whether or not a particular example was contained in the model's training dataset.
17. The loss function is a mathematical process that quantifies during the training the error between a model's prediction and the actual target value.

edps.europa.eu

QT-AE-24-001-EN-N

Publications Office
of the European Union

ISBN 978-92-9242-886-0
doi:10.2804/797140