TechSonar

2023-2024 REPORT

EDPS
European Data Protection Supervisor

Project Number: 2023.5263
Title: TechSonar 2023-2024 report

PDF/Volume_01	QT-02-23-189-EN-N	978-92-9242-820-4	10.2804/969613
PRINTED/Volume_01	QT-02-23-189-EN-C	978-92-9242-819-8	10.2804/924082
eBook/Volume_01	QT-02-23-189-EN-E	978-92-9242-821-1	10.2804/594449
Electronic Files/Volume_01	QT-02-23-189-EN-Q	978-92-9242-822-8	10.2804/36558

Readiness and adaptability in an evolving technology landscape

By Wojciech Wiewiórowski

The EDPS launched TechSonar in 2021, serving as its first foresight project. The rapidly evolving technological landscape requires us to anticipate new technological challenges, to be able to influence their evolution and use. A clear signal in this direction is the increasing pace of deployment of artificial intelligence in everyday life and machine learning applications, which requires a more proactive and anticipatory attitude towards an appropriate and effective governance of technology.

TechSonar is a process that empowers the EDPS to continuously analyse the technology arena with the aim of selecting tech trends we foresee for the following years. We have always been aware of the need to integrate the technological element within the data protection assessment process. EDPS initiatives, such as the creation in 2014 of the Internet Privacy Engineering Network (IPEN), exemplify this school of thought. However, the need to become proactive in our relationship with technology has become increasingly compelling and has led us to start our foresight journey.

Although we cannot fully predict the pace and direction of how technology will evolve, we can inform ourselves on how to forecast and prepare for possible outcomes and scenarios. There is a need for new tools and skills that will enable data protection authorities to intervene effectively and in a timely manner. The EDPS foresight effort follows a risk-based approach, focusing on technologies that are more likely to affect or harm individuals' rights to data protection and privacy.

My aforementioned remarks are especially true for rapid developments in the field of artificial intelligence. Currently, its use is increasing, posing ethical, legal, and technical questions in the field of fundamental rights, including data protection and privacy. AI technology has the potential to improve our lives and our safety and security, but it should never come at the cost of our dignity and values.

As a supervisory authority, we are convinced that taking active steps in the field of foresight will improve our way of working, ultimately establishing a continuous process from the identification of technology trends to the development and management of structured internal knowledge. This knowledge will then feed our advisory, supervisory and awareness-raising activities.

The Global Privacy Assembly (GPA) is a forum connecting over 130 data protection and privacy authorities. A few weeks ago, our TechSonar project was awarded the GPA Global Privacy and Data Protection Awards 2023 in the innovation category. The prize rewards the EDPS for forward-thinking and adaptive measures in response to disruptive technological models. We are grateful and proud that the GPA has recognised our efforts in this field.

The award reinforces the EDPS' commitment to remaining at the forefront of data protection, ensuring both preparedness and adaptability in a constantly evolving technological landscape. It calls for the need to reinforce the role of anticipatory and foresight techniques into our data protection activities and to support the value-creation process of privacy enhancing technologies. We want to carry out this commitment seeking synergies and collaboration with other data protection authorities and organisations that have undertaken a similar anticipatory approach.

Our new latest TechSonar report covers the topics of Large Language Models, Digital Identity Wallets, Internet of Behaviours, Extended Reality, and Deepfake Detection. We believe that these technologies deserve to be assessed in order to anticipate the potential positive and negative impacts of their future use and, where possible, to intervene.

Tech trends 2023-2024

Large Language Models (LLM)
pag.2

Digital Identity Wallet
pag.6

Internet of Behaviours (IoB)
pag.10

Extended Reality (XR)
pag.13

Deepfake Detection
pag.17

Large Language Models (LLM)

Author: Xabier Lareo

Language models are artificial intelligence (AI) systems designed to learn grammar, syntax and semantics of one or more languages to generate coherent and context-relevant language. Language models have been developed using neural networks since the 1990s, but the results were modest.

The evolution to large language models (LLMs) was made possible by technical developments that improved the performance and efficiency of AI systems. These developments included the advent of large-scale pre-trained models, the development of transformers (which learn context and meaning by tracking relationships in sequential data), and self-attention mechanisms (which allow models to weigh the importance of different elements in an input sequence and dynamically adjust their influence on the output).

As a type of generative AI system, LLMs create new content in response to user commands based on their training data. They are trained on huge amounts of text sources (from millions to billions of words) from a variety of sources, including public sources, and their size can be measured by the number of parameters used.

They're also considered a type of 'foundation model', which is a model trained on large amounts of data (usually using large-scale self-monitoring) that can be adapted to a variety of applications, including text generation, summarising, translating, answering questions, and more.

The number of parameters in LLMs has increased over time: while version 2 of the Generative Pre-trained Transformer (GPT-2) had 1.5 billion parameters, the Pathways Language Model (PaLM) reached 540 billion parameters.

At a certain point, the development of competitive high-performance LLMs seemed to be something that only the most resourceful technology companies, such as Google, Meta or OpenAI, could achieve. However, two developments changed that trend and made LLM development more broadly available. First, the publication of research showing that there is an optimal set of values when selecting computing power, model size and training dataset size. Second, the appearance of parameter efficient fine-tuning techniques (e.g. LoRA), which have greatly reduced the amount of resources needed to train an LLM - PALM 2 already follows this trend and, although it appears to have been trained with a much larger dataset, it has fewer parameters than its predecessor (340 billion against PaLM's 540 billion).

Some LLM service providers have made their models publicly available – after registration and, in several cases, requiring a subscription model - through web interfaces that allow users to enter commands (prompts) and view the output generated by the models. Publicly accessible models are sometimes presented as research previews or testing versions that might produce erroneous or harmful output. LLM service providers also tend to offer access to their models (usually for a fee) through an application programming interface (API) that allows their LLM to be embedded into customers' IT systems.

LLMs are currently being used or tested for a wide variety of tasks in different domains, including translation; customer care (e.g. chatbots); education (e.g. language training); natural language processing (e.g. named entity recognition or summarisation); supporting the generation of images from a given prompt output; preparation of programming code; or even the creation of artistic works.

As LLMs continue to evolve, they both offer opportunities and important challenges for privacy and data protection.

Positive impacts foreseen on data protection

LLMs could be used to support certain privacy activities in very specific scenarios, if designed, developed and deployed in a responsible and trustworthy manner, respecting the principles of data protection, privacy, human control and transparency.
For example:

- **Detection of personal data**
  Identifying personal data in unstructured data, such as in text fields is relatively

easy for humans, but difficult to automate using simple rules. However, human review does not scale well and becomes impractical or unfeasible in large-text files or web-scraped datasets. The natural language processing capabilities of LLMs could help detect and better manage personal data on unstructured information (e.g. a text field containing family history). LLMs could also help reduce the personal data included in their training datasets, by automatically identifying, redacting or obfuscating personal data.

Negative impacts foreseen on data protection

- **Training LLMs is a data-intensive activity, which can include personal data**
  The vast majority of the data used to train state-of-the-art LLMs are texts scraped from publicly available Internet resources (e.g. the latest Common Crawl dataset, which contains data from more than 3 billion pages). These web-scraped datasets contain personal data of public figures, but also of other individuals. Personal data contained in these datasets could be accurate or inaccurate. These datasets could also contain plain misinformation. Implementing controls to address the data protection risks posed by the use of these datasets is very challenging. Moreover, if not properly secured, LLM output might reveal sensitive or private information included in the datasets used for training, leading to potential or real data breaches.

- **"Hallucinations", data accuracy and bias**
  LLMs sometimes suffer from so-called 'hallucinations', meaning they produce erroneous information that appears to be correct. When hallucinating, an LLM can produce false or misleading information about individuals. Inaccurate information can affect individuals not only because it can damage their public image, but also because it can lead to decisions that affect them. LLMs, if trained on biased data, could perpetuate or even amplify biases present in their training data. This might lead to unfair or discriminatory outputs, potentially violating the principle of fair processing of personal data.

- **Implementing data subjects' rights is difficult**
  LLMs store the data they learn in the form of the value of billions or trillions of parameters, rather than in a traditional database. For this reason, rectifying, deleting or even requesting access to personal data learned by LLMs, whether it is accurate or made up of "hallucinations", may be difficult or impossible.

Suggestions for further reading:

- Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. "Attention is All you Need", 2017, **https://doi.org/10.48550/arXiv.1706.03762**
- Kaplan, Jared, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu and Dario Amodei. "Scaling Laws for Neural Language Models", 2020, **https://doi.org/10.48550/arXiv.2001.08361**
- Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "LoRA: Low-rank adaptation of large language models", 2021, **https://arxiv.org/abs/2106.09685v2**
- Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. "A comprehensive overview of large language models" , 2023, **https://doi.org/10.48550/arXiv.2307.06435**
- Global Privacy Assembly Resolution on Generative Artificial Intelligence Systems, 2023, **https://edps.europa.eu/system/files/2023-10/edps-gpa-resolution-on-generative-ai-systems_en.pdf**

Digital Identity Wallet

Author: Massimo Attoresi

A Digital Identity Wallet (DIW) is an application that allows the secure storage, management, and sharing of personal identification data, credentials and other pieces of information, often called "attributes", relating to the owner of that virtual wallet. Think of it as a digital version of your physical wallet, but instead of holding tangible items like cash or credit cards, it holds digital attributes. Digital identity wallets can exist in various forms, including mobile apps, browser extensions, or even dedicated hardware devices.

DIW content can vary from unique alphanumeric identifiers and natural identification data, such as first and second name, address, birth date and place, to elements such as driving licences, credentials to access places (e.g. a sport facility) and resources (e.g. public transports), certifications, debit/credit cards etc., including digital currencies. A DIW can potentially contain any kind of digital content related to that individual. In that regard, DIW can feature functionalities similar to those of a Personal Information Management System.

Parties that guarantee their authenticity and integrity usually release DIW attributes. For example, an accredited authority may issue a professional certificate, a competent public administration may issue a driving license, or a library may issue credentials allowing people to access its resources and borrow items.

In a nutshell, a DIW can be used to identify and authenticate an individual or to authorise that individual to access a resource against the same party that issued the attributes or against a third party (also called "relying party").

The trustworthy nature of the attributes released by issuers is usually ensured by the use of a cryptographic "signature" derived from a hierarchy of commonly trusted third parties, traditionally "certification authorities". Other DIW schemes exist with various technical and governance architectures. For example, "self-sovereign identities" based schemes exist that leverage decentralised identifiers, which are trustworthy globally unique identifiers directly generated and controlled by an individual or organisation.

They are used for identification or authentication against other individuals or organisations (e.g. a service provider) without the need of identity service providers, and certificate authorities.

There is a wide spectrum of projects for the use of DIW in the public and private spheres. This is due also to legislative initiatives such as for cross-border electronic identification, authorisation and trust services in the EU (eIDAS), where DIW are planned to foster a variety of use cases, including a possible digital euro currency.

Positive impacts foreseen on data protection

- **Increase of confidentiality and integrity of personal data**
  All pieces of information within a DIW are to be provided by their sources with a proof of origin, thus ensuring authenticity (which combines confidentiality and integrity) to any party relying on that information. For example, the natural identifiers of a person can be guaranteed by the civil registry for any third party that requires them.

- **Increase of personal data accuracy**
  Based on the proof of origin and information integrity safeguards, DIWs can give higher assurance that the pieces of information relating to the owner are accurate and up-to-date. For example, the amount and type of social benefits stored in a DIW can be guaranteed by the public administration issuing those benefits and legitimately updated when necessary. Similarly, individuals will be able to keep up-to-date information on themselves such as interests and preferences, to be directly collected from the DIW and thus always under the user responsibility.

- **Enhanced control for data subjects**
  In principle, yet depending on the implementation, individuals could be more in control of the data stored in their

DIW. Trustworthy personal information can be securely accessed directly in the DIW, based on user's preferences (when there is no obligation by law). This would avoid unnecessary dissemination in databases of the relying parties. Furthermore, even in circumstances when they are not the providers of the information stored in their DIW, individuals could always be aware of their personal data and of who has access to them.

Negative impacts foreseen on data protection

- **Increased risk of profiling**
  DIWs intrinsically carry individuals' identification information as well as other pieces of information that could uniquely identify them. In absence of safeguards, this information could be combined by all parties having access to the DIW (providers of identity services in particular but also relying parties) with other information already retained by those parties on the actions performed by the same individual. Furthermore, DIWs can store any possible personal data including sensitive ones, directly or indirectly relating to health, sexual orientation, religious or philosophical beliefs, political opinions, financial situation, family life, etc. This accumulation of personal information could encourage both private and public actors' appetite to exploit this data. For this reason, DIWs have a high potential to enable profiling of individuals if the features and use of DIWs are not consistent with a privacy by design and by default approach, and if appropriate policies are not in place. Some specific weaknesses enabling profiling are described below.

- **Unnecessary/disproportionate disclosure of personal data**
  Depending on the implementation, there is a risk that providers of identity services and relying parties access more pieces of information stored in DIWs than what they really are allowed to, based on individuals' consent or other lawful bases. This can be due to an inadequate policy or design choice, neglecting data minimisation requirements.

- **No data minimisation: abuse of identification instead of authorisation**
  In certain use cases, it is necessary to identify/authenticate the individuals unambiguously to be able to relate to that individual. The law usually provides for these circumstances. In other use cases, it is only necessary to demonstrate that a specific individual is authorised to access a specific resource, yet it is inappropriate common practice to disclose identification data to that purpose. For example, once registered, to access a library it is sufficient to produce an authorisation, it is not necessary to disclose your identity.

Suggestions for further reading:

- European Commission, Shaping Europe’s Digital future, **Discover eIDAS**
- BEUC, **Making European Digital Identity as safe as it is needed - BEUC position pape**r, 10 February 2022
- ENISA, **Digital Identity: Leveraging the SSI Concept to Build Trust**, 20 January 2022

EDPS related work:

- EDPS, **IPEN Workshop on Digital Identity**, 22 June 2002
- EDPS, **Formal comments of the EDPS on the Proposal for a Regulation of the European Parliament and of the Council amending Regulation (EU) No 910/2014 as regards establishing a framework for a European Digital Identity, 28 July 2021**
- EDPS, **TechDispatch #3/2020 - Personal Information Management Systems**

Internet of Behaviours (IoB)

Author: Xabier Lareo

In 2012, Professor Gote Nyman coined the term **Internet of Behaviours** (IoB) to describe a network in which behavioural patterns would have an IoB address in the same way that each device has an IP address in the Internet of Things (IoT).

However, the term IoB is most often used to describe an extension of the Internet of Things (IoT). A network of interconnected physical and digital objects that collect and exchange information over the Internet, linking this data to specific human measured or inferred behaviours. This is referred to as "General IoB". Gartner Consulting highlighted the Internet of Behaviours as one of the **Top Strategic Technology Trends for 2021**.

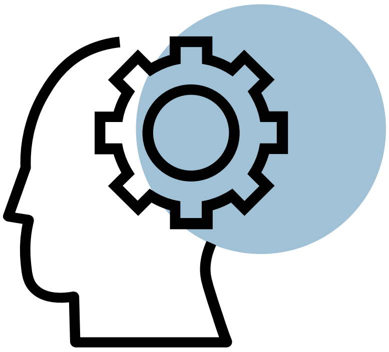
The aim of IoB is to address how data collected can be interpreted from a human psychological and sociological perspective and how to use this understanding to influence or change human behaviour for various purposes, ranging from commercial interests to public policies.

Overall, IoB is not a completely new concept: behavioural-targeted advertising tracks human behaviour to show personalised ads, or Bluetooth and Wi-Fi technologies are used in malls to infer prospective shoppers' behaviours with a view to better marketing. IoB somehow integrates extensively all these technologies in a holistic approach, and are able to follow people's lives and behaviours whenever it is possible to measure their interaction with the digital or physical objects surrounding or interacting with them.

An example of an application of IoB could be the use of patients' and employees' location data in hospitals during the COVID-19 pandemic to identify the behaviours that spread or mitigate the virus, in order to be able to influence future ones for people's benefit (e.g. RFID tags at handwashing stations to identify if employees are following hygiene protocols). Information from RFID readers could be used to track when and how often healthcare workers or patients are washing their hands and place reminding messages in relevant spots. Computer vision could detect non-compliance with preventive policies, such as the obligation of wearing masks, and trigger reminders on the closest screen.

Negative impacts foreseen on data protection

- **Increased processing of personal data and profiling**
  General IoB relies on the collection and processing of data from different IoT devices, such as wearables, smart cameras or Bluetooth and Wi-Fi sensors. These devices include identifiers (e.g. IP, MAC or email addresses) that make it possible to cross-link, profile and identify individuals. This increased processing of personal data - possibly by different actors and for different purposes - might easily conflict with the principles of data minimisation and purpose limitation.



- **Lack of transparency and control**
  IoT devices suffer from transparency and control issues because they often lack appropriate means to inform their users (e.g. tiny screens or absence of it or of any other form of notice), their data collection is seamless (e.g. surveillance cameras) and the means to exert control over the processing are limited. The General IoB inherits these transparency issues and could make them even greater if IoB users are not properly informed of the way their behavioural data are processed.

- **Potential inaccuracy**
  The General IoB works on the assumption that human behaviour can be accurately inferred by tracking individuals. However, this might not be the case in many contexts, due to weaknesses inherited from technologies used for inferences, such as Machine Learning (e.g. bias), and the complexity of the link between human behaviours and the rationale behind them. Data controllers will hardly be able to ensure data accuracy unless they clearly inform the individuals subject to the IoB and provide them with the means to rectify erroneous inferences.

Suggestions for further reading:

- Sun, Jiayi, Wensheng Gan, Han-Chieh Chao, S. Yu Philip, and Weiping Ding. "Internet of Behaviors: A survey", 2023, **https://doi.org/10.48550/arXiv.2211.15588**
- Moghaddam, Mahyar T., Henry Muccini, Julie Dugdale, and Mikkel Baun Kjægaard, "Designing internet of behaviors systems", 2022, **https://doi.org/10.48550/arXiv.2201.02022**
- Tariq Rahaman, "Smart Things are Getting Smarter: An Introduction to the Internet of Behaviors", Nova Southeaster University, 2022, **https://doi.org/10.1080/02763869.2022.2021046**

Extended Reality (XR)

Author: Vítor Bernardo

Extended reality (XR) is an emerging umbrella term for all immersive technologies, including virtual reality, augmented reality, and mixed reality.

Virtual reality (VR) is a technology that creates a digitally simulated immersive environment or experience for users. It typically involves the use of specialised hardware and software to create a computer-generated 3D spatial, and possibly multi-sensorial, environment that can be interacted with in a seemingly real or physical way. VR allows users to experience and interact with a digital environment as though they were real.

Augmented reality (AR) on the other hand is a technology that overlays digital information, such as text; images; sound; videos; or 3D models, onto the real-world environment. AR enhances the real world by adding computer-generated elements to it.

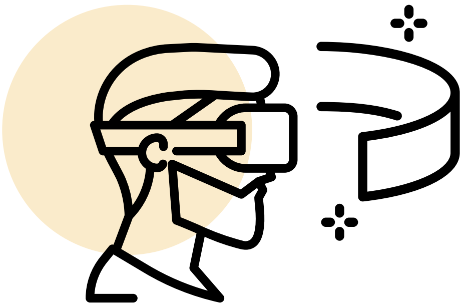
Mixed reality (MR) systems are immersive technologies that bring physical objects into digital environments or digital objects into physical reality. One type of MR is Cinematic Reality, offering immersive 360 degrees viewing with live camera footage.

XR technologies typically rely on smartphones, tablets, smart glasses, or other wearable devices to deliver the augmented and/or virtual experiences. The wearables are also required to collect certain basic information provided by the user as a starting point, and then a continuous stream of new feedback data generated as the user interacts with their virtual environments to create the illusion of interaction with the virtual elements.

Whether combined or alone, these technologies can have many applications. Professional training, entertainment, education, and architecture are some of the fields that are expected to be changed profoundly as VR evolves. AR is expected to provide users with valuable context-related information on the real world, enhancing their understanding of the environment.
These technologies can have many benefits in several different fields. They can provide contextual information during surgery in healthcare, information about sights or museums and historical augmented environments in tourism, and directions and/or warnings in navigation.

Positive impacts foreseen on data protection

- **Providing information to data subjects with AR**
  Augmented reality's ability to provide contextual information (i.e. available in a specific area or in the presence of a specific object) can also be used to provide more information about the processing of personal data. For example, individuals entering a CCTV-covered area with AR-enabled devices could be presented with information about the data controller, the purpose of the data processing, and possible ways to exercise their rights.
  It should be borne in mind that, in such situations, AR would be an additional channel to provide information that should not replace the mechanisms already in place. In addition, the use of these mechanisms for the provision of information should not lead to further processing of the data for other purposes.



Negative impacts foreseen on data protection

- **Intensive collection of personal data from users and user profiling**
  VR systems might capture user's behaviours, such as head orientation, and position. Some systems can track other body part movements to increase immersion (e.g. hand, feet, chest, elbow or knee). XR can also incorporate gaze tracking, respiration, heart rate or even brain-computer interface (BCI) neural signal interpretation. User movement data can be collected at frequencies of up to 1000Hz, meaning that systems can take 1000 user measurements per second.
  This can lead to intensive data collection of multiple characteristics from users that can allow the definition of a detailed description of characteristics and behaviours. Not only can these data collections contain multiple types of personal and possibly sensitive information, but XR devices can also combine this information to reveal or infer additional details about individual

users (distance from the floor, for instance, can be used to infer the user's height). According to some authors, head position and movement can be used to infer neurological conditions such as attention deficit hyperactivity disorder, autism or dementia. Finally, most VR services currently on the market require users to log in to the device, further increasing the risk of user profiling across different devices.

Such a high volume of data processing is difficult to reconcile with the principles of data minimising and purpose limitation.

- **Unintentional disclosure of personal data**
  In VR, an avatar is a digital representation of a user or player within the virtual environment. Avatars can be customised to varying degrees, depending on the VR system or platform. Users have the ability to choose different appearances, clothing, and even gestures or expressions to personalise their virtual identities.
  However, studies have shown that when configuring avatars for social VR, people tend to construct avatars that match their physical selves, reflecting their aesthetics, gender, ethnicity and age/maturity, increasing the risk of identification. There is also research indicating the possibility to correlate the movements of an avatar with data of users' movement recorded whilst performing a set of movements in real life.

  Users can become emotionally immersed in these virtual spaces, which they may be able to access through different devices using a single virtual identity. This may make them more likely to (unintentionally) disclose personal information in the immersive environment that they would not otherwise.

  Additionally, VR environments are evolving into complete virtual worlds with the persistence of the user's online actions over time (i.e. the state of the user's actions is preserved in the VR environment to replicate the physics of the real world). In such a scenario, there would be an increased risk of revealing personal data from the user's online activities (e.g. activities reflecting political views or personal interests).

- **Personal data collection from non-users**
  AR systems are designed to interact with the user's real environment. When in use, they can continuously collect video and sound from the user's surroundings, which may include other people who are not users and who are unaware of the data processing. This can lead to unauthorised collection of data from others.
  Unauthorised data collection from the real environment may also occur in the case of VR, especially since some of the largest social media providers have shown interest in creating virtual reality environments that duplicate the real world we live in (i.e. an environment that replicates the physical objects and beings of the real world). Although still a concept, such an endeavour would require a massive collection of unauthorised personal data.

Suggestions for further reading:

- Dick, E. (2021). *Balancing user privacy and innovation in augmented and virtual reality Information Technology and Innovation Foundation*
- Miller, M. R., Herrera, F., Jun, H., Landay, J. A., & Bailenson, J. N. (2020). *Personal identifiability of user tracking data during observation of 360-degree VR video Scientific Reports*, 10(1), 17404
- Roesner, F., Kohno, T., & Molnar, D. (2014). *Security and privacy for augmented reality systems Communications of the ACM*, 57(4), 88-96

Deepfake Detection

Author: Vítor Bernardo

A deepfake is the manipulation or artificial generation (synthesis) of audio, video or other forms of digital content to make it appear that a particular event occurred, or that someone behaved or looked differently than they actually did.
The manipulation of photos and videos, which used to be done manually using graphical editing tools, has undergone a significant evolution through the use of artificial intelligence and, in particular, deep learning.

Among the various deepfake creation methods, Generative Adversarial Network (GAN) is a technology that has shown remarkable results, creating manipulations that are difficult to distinguish from original content. GANs are machine learning (ML) models in which two neural networks - a generator and a discriminator - compete with each other to make predictions that are as accurate as possible or, in the case of deepfake generation, to produce the most realistic result.

In addition to the question of content manipulation, there is also the concern that deepfake content could promote disinformation and have a negative impact on people’s opinions, with potential political and social consequences. Nude or otherwise offensive depictions of people, hoaxes and financial fraud can also be produced through

video manipulation.
Additionally, the ability to impersonate other people, by swapping faces in photos and videos, increases the risk of unauthorised access to services or premises.

Several approaches have been proposed to automatically detect fake videos of people, including eyebrow change detection, eye blink and movement detection, inconsistent corneal specular highlights (i.e., consistency in the eye's reflection of ambient lighting), and even heartbeat detection by capturing slight skin colour changes in the video.
Other techniques have focused on the detection of unique elements (fingerprints) in the digital content resulting from the use of deepfake tools; these elements are commonly referred to as 'artifacts'.
Categorisation algorithms are trained on large collections of real and fake audio-visual samples to identify *artifacts*.

The existing deepfake detectors rely mainly on the signatures of existing deepfake content by using ML techniques, including unsupervised clustering and supervised classification methods, and therefore are less likely to detect unknown deepfake manipulations. However, the technology used for deepfake detection content is still not able to provide sufficient assurance. The current deepfake detectors face challenges, particularly due to incomplete, sparse, and noisy data in training phases.

Positive impacts foreseen on data protection

- **Prevention of the impact of deepfakes on individuals**
  With the limitations noted above, deepfake detection can be used to identify content that has been manipulated for malicious purposes. Detecting and tagging fake videos and images allow individuals and organisations to take action to stop the spread of potentially damaging misinformation. This can safeguard the reputation and privacy of individuals and prevent the dissemination of fake news, frauds, or cyberbullying.

- **Protection of personal data by preventing deepfake-based attacks**
  Deepfake manipulations can be used to create convincing impersonations of individuals, potentially leading to identity theft or unauthorised access to sensitive data. As fake videos and audio are used in various forms of cyberattacks, including spear phishing and social engineering, having robust detection mechanisms in place can prevent unauthorized access to sensitive information.

- **Improvement of data accuracy by applying data validation**
  Deepfake detection can be used for data validation. In the financial, healthcare, and legal sectors, are examples where data accuracy is paramount, deepfake detection tools can help verify the authenticity of

of documents, audio recordings, or video footage, ensuring that decisions and actions are based on reliable information.

Negative impacts foreseen on data protection

- **Lack of fairness and trust**
  Research indicates that data bases commonly used for training of deepfake detection lack diversity and, more importantly, show that deepfake detection models can be strongly biased. It has been observed that existing audio and visual deepfake datasets contain imbalanced data of different ethnic origins and genders. In some situations, having large lips or nose, being heavier or black led to more detection errors compared to images without these attributes. There is a risk that the application of biased models in the real world could discriminate against certain individuals.

- **Lack of transparency and fairness in detection methods**
  Existing deepfake detection approaches are typically designed to perform batch analysis over a large dataset. However, when these techniques are employed in the field, for example by journalists or law enforcement, there may only be a small set of videos available for analysis.
  In these situations, an explanation of the numerical score given to the likelihood of the content being deepfake may be necessary for the analysis to be trusted before publication or utilization in possible legal actions. However, most deepfake detection methods and tools lack such an explanation, especially those based on deep learning, due to their black-box nature.

- **Lack of accuracy**
  Presently, deepfake detection methods are formulated as a binary classification problem, where each sample can be either real or fake. However, for real-world scenarios, videos can be altered in ways other than deepfake (for instance, by post-production), so content not detected as manipulated does not guarantee that the video is a genuine one. Additionally, fake images and videos are usually shared on social networks and for this reason suffer from high variations, such as compression level, resizing, and noise (a process known as media washing). This can incur in a large number of false negatives (i.e., undetected fakes).

Suggestions for further reading:

- Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4), 3974-4026
- Patil, K., Kale, S., Dhokey, J., & Gulhane, A. (2023). Deepfake detection using biological features: a survey. arXiv preprint arXiv:2301.05819
- Trinh, L., & Liu, Y. (2021). An examination of fairness of AI models for deepfake detection. *arXiv preprint arXiv:2105.00558*

edps.europa.eu

@EU_EDPS

EDPS

European Data Protection Supervisor

@EDPS@social.network.europa.eu

@EDPS@tube.network.europa.eu

QT-02-23-189-EN-N

Publications Office
of the European Union

ISBN 978-92-9242-820-4
doi:10.2804/969613