Cybersecurity challenges in Artificial Intelligence (AI)

Insights and recommendations from Cyberwatching.eu

2021

cyberwatching.eu consortium

Trust-IT Services
Communicating ICT to markets

Oxford e-Research Centre

University of Oxford

ICT Legal Consulting
Balboni Bolognini & Partners

European Digital SME Alliance

CONCEPTIVITY

360 SECURITY

AON

aeiciberseguridad
Agrupación Empresarial Innovadora CIBERSEGURIDAD y Tecnologías Avanzadas

Table of content

Disclaimer

The work described in this document has been conducted within the project cyberwatching.eu. This project has received funding from the European Union's Horizon 2020 (H2020) research and innovation programme under Grant Agreement no.740129. This document does not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of its content.

1 Introduction

Artificial intelligence (AI) is increasingly becoming an integral part of technology and cyberspace. AI can be implemented in systems, software and devices of varying sectors, to similar degrees of effectiveness.[1] From the data protection perspective, AI is typically used as a tool for automated decision-making and profiling, by leveraging algorithms to process large volumes of data.[2] In terms of AI being implemented in critical infrastructures, countries are putting AI to use in order to offer better and faster telecommunication services to citizens, run trade and stock markets by algorithms, or even create governmental procedures for voting, and managing administrative complaints.[3] In this context, **the main challenges arise when the processing activities carried out by means of AI are capable of leading to automated decisions** which produce legal, or similarly significant effects on data subjects.[4]

2 Challenges of Data Minimisation

One of the typical assumptions around the use of AI is that a large (potentially, a progressively expanding) dataset will be needed, so that the AI's algorithm can generate accurate and useful results, or even further develop (in the case of machine-learning algorithms). Considering that such large datasets may also include personal data, questions immediately come to mind: **Is it feasible for an AI-based system to work effectively without resorting to large volumes of (personal) data?** How can the principle of data minimisation be adhered to by AI-based systems?

This challenge is at the heart of the data protection issues that arise from the use of AI, because it seemingly places a core GDPR principle against the purposes and functions of AI itself. The incentive is to collect as much data as possible in order to render the AI operational, which may not factor in any considerations for the principle of data minimisation[5] – particularly because **AI algorithms will not only rely on data which is strictly relevant** to reach a desired output, given that AI must also learn to identify and discard data which is irrelevant to that goal, in order to increase its effectiveness after deployment (and avoid inaccurate or discriminatory results).[6] This shows that even contextual data can be important for AI.[7] Therefore, the requirements for data minimisation – processing only personal data which is "*adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed*",[8] cannot be upheld in the traditional form of the GDPR. Furthermore, given the difficulties in predicting the training speed or accuracy of a given AI model, AI developers may not be capable of correctly predicting just how much data (necessary or contextual) an AI algorithm will need in order to deliver the expected output.

In order to tackle this concern, it would be **recommended that further research be carried out on how the concept of data minimization can be tackled when mass data collection is necessary,**

[1] For more on this, see Consultative Committee of the Convention of the Protection of Individuals with regard to Automatic Processing of Personal Data, *Guidelines on Artificial Intelligence and Data Protection* (25 January 2019), available at: https://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8.

[2] For more on this, see UK Information Commissioner's Office, Big *data, artificial intelligence, machine learning and data protection* (4 September 2017), available at: https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

[3] European Commission, *Statement on Artificial Intelligence, Robotics and* 'Autonomous' *Systems* (9 March 2018), available at: https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

[4] See Art. 22(1) GDPR.

[5] Please note that this issue arises typically only for long-term AI projects, which integrate the data collected during the service or product lifecycle. If an AI is being developed as a one-off exercise, then once the training of the algorithm has occurred, there is no longer a conflict with the data minimization principle (as the storage and processing of large training datasets is no longer required).

[6] For more information on this, see European Parliament, Understanding *algorithmic decision-making: Opportunities and challenges* (March 2019), available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf.

[7] For simplicity's sake, consider the following example: if a developer is building an AI-based system to visually recognize fruit, the AI's training dataset may also need to include not only images of fruit, but also of any other objects or materials that may be mistaken for fruit, so that the AI learns what input to reject (and not just what input to accept).

[8] Art. 5(1)(c) GDPR.

in order to train algorithms within AI models. It appears relevant, in this context, to distinguish between the data minimization during (1) the training of the model algorithm (original training data), and (2) once the AI is actually running on real-time data, which would be needed to ensure fairness, accuracy and lack of discrimination in AI decision-making. Furthermore, when assessing the adequacy, necessity or relevance of a given dataset for AI-based processing activities, due consideration should be given to the complexity of the problem or processing that the AI model is targeting, as well as the complexity of the learning algorithm. Considering the complexity of the problem/processing in question can help **define the underlying functions which the algorithm is meant to achieve**, providing insight into the input variables (i.e., types and volumes of data) that the algorithm will require; considering the complexity of the learning algorithm can help to understand how such data will be 'parsed' through the algorithm, allowing for a more precise identification of types of (personal) data which could be considered as adequate, relevant and necessary for the AI model to meet its intended purpose.

3 Challenges of Purpose Limitation

Under the GDPR's principle of purpose limitation, [9] personal data must be "*collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes*".[10] As noted by the Article 29 Data Protection Working Party, "*any processing following collection, whether for the purposes initially specified or for any additional purposes, must be considered 'further processing' and must thus meet the requirement of compatibility*."[11] This notion of 'compatibility' is further explored in Art. 6(4) GDPR, which lays down criteria to be assessed by a controller in order to establish if a further processing purpose is compatible with the initial purpose for data collection:[12] (1) whether there is any link between these purposes; (2) the context in which the personal data was collected;[13] (3) the nature of the personal data in question;[14] (4) the possible consequences of the intended further processing for data subjects; and (5) the existence of appropriate safeguards, such as encryption or pseudonymisation.

In light of this, a separate issue, which may arise more commonly with **machine-learning algorithms**, is the possibility for such algorithms to, **autonomously** (and in unexpected or unpredictable ways) **process (personal) data for purposes different**, or incompatible with, the original purposes for which the algorithms were set up. Machine-learning-based algorithms can - not only learn to achieve the goals they are programmed for - but they can also reinterpret their goals, shifting the focus from achieving their original goals to achieving the feedback they would receive if they had done so.[15] Where this occurs the result is that **personal data is processed for a purpose not originally disclosed to data subjects** (i.e., not specified or explicit), and which may potentially be incompatible with the purposes for which personal data was originally collected. Such a result would inevitably collide with the principle of purpose limitation.

Such a concern can seemingly only be addressed by **imposing limitations or further requirements on the use of personal data within AI-based systems.** Algorithms (and, in particular, machine-learning algorithms) should be carefully developed so that they will not, autonomously or beyond the control of the relevant controller, process personal data collected for purposes beyond the scope of their collection (or, at least, not without a proper compatibility test, under Art. 6(4) GDPR, having been performed by the relevant controller) – **any guidance which can be offered by policy-makers and competent authorities in this regard would prove invaluable.**

Controllers should carefully analyse the systems that they wish to implement and ensure that they are able to provide clear and adequate information to data subjects on how those systems will work and, in particular, the purposes for which they will use personal data – here, **guidelines or templates on how to disclose such information in a digestible way for individuals (consumers), considering, where relevant, the requirements of Art. 13(2)(f) and 14(2)(g) GDPR,[16] could be of great benefit to AI developers and users.**

[9] Art. 5(1)(b) GDPR.

[10] On this point, see Article 29 Data Protection Working Party, *Opinion 03/2013 on purpose limitation* (2 April 2013, available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf), p. 15: "*Personal data must be collected for specified purposes. The controller must therefore carefully consider what purpose or purposes the personal data will be used for, and must not collect personal data which are not necessary, adequate or relevant for the purpose or purposes which are intended to be served*".

[11] Article 29 Data Protection Working Party, *Opinion 03/*2013 *on purpose limitation* (2 April 2013), p. 21, available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

[12] Note that Art. 6(4) GDPR generally allows further processing to take place, even in the absence of compatibility with the original processing purposes, where consent is relied on as a legal basis for the further processing, or where the further processing is authorised by Union or Member State law.

[13] Under Art. 6(4)(b) GDPR, the relationship between data subjects and the controller must be considered, in particular.

[14] Under Art. 6(4)(c) GDPR, whether or not special categories of personal data, or personal data related to criminal convictions and offences, are processed is an important consideration in this regard.

[15] For more on this, see, e.g., Casey Chu et al, CycleGAN, a Master of Steganography, available at: https://arxiv.org/pdf/1712.02950.pdf.

[16] See Section 3.1.3, below.

4 Challenges of Transparency and Lawfulness

Under the GDPR's principle of transparency,[17] **controllers are required to provide data subjects with information as to their activities involving the processing of personal data**, under, e.g., Arts. 13 and 14 GDPR. This information must include, in particular and where automated decision-making is concerned, "*the existence of automated decision-making, including profiling, (...) and (...) meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject*".[18] This seeks to ensure that data subjects understand exactly how their personal data will be used by a given controller, and what the consequences for them may be.

When AI-based systems are used to process personal data, **difficulties arise in the provision of clear information to data subjects**, not only because such systems are often very complex (and, so, hard to explain in a concise and intelligible manner to data subjects, as required by Art. 12(1) GDPR), but also because the purposes for which such systems may handle personal data may evolve over time.[19]

According to the European Commission High-Level Expert Group on AI (AI HLEG), the requirement of transparency in AI "*is closely linked with the principle of explicability and encompasses transparency of elements relevant to an AI system: the data, the system and the business models*".[20] As the Ethics Guidelines for Trustworthy Artificial Intelligence establish, traceability,[21] explainability[22] and communication[23] all play fundamental roles in transparency.

One particular specification of this issue, which involves also the principle of lawfulness, is the **selection of an appropriate legal basis for the use of AI**. As noted above,[24] controllers wishing to use AI to carry out automated individual decision-making will not only have to identify a legal basis, under Art. 6 GDPR, but also ensure that an exception, under Art. 22(2) GDPR, applies to their specific case. In particular, in the absence of Union or Member State law authorising the use of AI

[17] Art. 5(1)(a) GDPR.

[18] Art. 13(2)(f) and 14(2)(g) GDPR. For more information on this, see Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (6 February 2018), available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053, and Article 29 Data Protection Working Party, *Guidelines on transparency under Regulation 2016/679* (11 April 2018), available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227.

[19] See Section 3.1.2, above. It should be noted that data subjects must be informed by controllers of the purposes for which personal data are to be processed, under Arts. 13(1)(c) and 14(1)(c) GDPR; this is also a result of the need for purposes to be explicit, under the principle of purpose limitation, reflected in Art. 5(1)(b) GDPR. For more on this, see Article 29 Data Protection Working Party, *Opinion 03/2013 on purpose limitation* (2 April 2013), available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf, and Article 29 Data Protection Working Party, *Guidelines on transparency under Regulation 2016/679* (11 April 2018), available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227.

[20] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for a trustworthy AI* (8 April 2019), pp. 18, 28-29, available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[21] Traceability, in this context, calls for the datasets that contribute to the AI's decision-making to be traceable, and that the algorithms used by the AI are adequately documented. This requires the establishment of procedures and methods that concretely ensure traceability, ensuring that all possible outcomes of the decisions made by the AI are known and traceable, as well as hypothetical decisions that the AI could make. See High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for a trustworthy AI* (8 April 2019), p. 18, available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[22] Explainability, on the other hand, requires an assessment of how decisions made by an AI are understood, how much AI-made decisions can affect its own decision-making processes, why the system was deployed and what the business model of the system is – in other words, AI-based systems must be designed in a manner which allows them to be explained to the individuals concerned. See High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for a trustworthy AI (8 April 2019),* p. 29, available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[23] Communication, the third requirement for transparency, entails the use of a disclaimer, allowing individuals to understand that they are interacting with an AI as opposed to a human being, also communicating the risks inherent to the AI. See High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for a trustworthy AI (8 April 2019),* p. 29, available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[24] See Section 2.1, above.

in this manner, controllers will be met with a choice: either Art. 22(2)(a) GDPR[25] is applicable, and therefore, they must rely on Art. 6(1)(b) GDPR,[26] or Art. 22(2)(c) GDPR[27] is applicable, and therefore, they must rely on explicit consent from the data subjects concerned.

However, both of these options represent particular challenges: Art. 6(1)(b) GDPR requires the processing in question to be objectively necessary for either the performance of a contract with a data subject, or to take pre-contractual steps at the data subject's request – if realistic and less intrusive options can be relied on to do so, this legal basis cannot be relied on;[28] consent, in turn, must be informed, which requires a minimum amount of information to be provided to data subjects about the processing to which they are consenting – naturally, if the processing purposes change, or other substantial parts of the information provided change, the validity of the consent itself may be called into question.[29] Outside of the scope of Art. 22 GDPR (such as where the decisions made do not create a legal or similarly significant effect on individuals, including, e.g., for the performance of analytics which are not used to make decisions on individuals,[30] or where there is substantial human intervention in an AI-based decision-making process[31]), controllers may consider other legal basis, including the pursuit of legitimate interests under Art. 6(1)(f) GDPR – this, however, will require a comprehensive legitimate interests assessment, as noted above.[32]

The challenges faced by AI in terms of transparency and lawfulness can be seen as sharing **similarities with the processing of personal data for scientific research purposes** – as noted by Recital 33 GDPR, *"[i]t is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research*". This Recital goes on to suggest that "*[d]ata subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose*".[33]

[25] Art. 22(2)(a) GDPR allows the processing of personal data in connection with automated individual decision-making if this is "*necessary for entering into, or performance of, a contract between the data subject and a data controller*".

[26] Art. 6(1)(b) GDPR allows the processing of personal data, in general, if this is "necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract".

[27] Art. 22(2)(c) GDPR allows the processing of personal data in connection with automated individual decision-making if this is "*based on the data subject's explicit consent*".

[28] European Data Protection Board, Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects (8 October 2019), p. 8, available at: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines-art_6-1-b-adopted_after_public_consultation_en.pdf.

[29] Article 29 Data Protection Working Party, Guidelines on consent under Regulation 2016/679 (10 April 2018, available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051), p. 18: "*(...) controllers do need to obtain a new and specific consent if purposes for data processing change after consent was obtained or if an additional purpose is envisaged*" and p. 21: "*There is no specific time limit in the GDPR for how long consent will last. How long consent lasts will depend on the context, the scope of the original consent and the expectations of the data subject. If the processing operations change or evolve considerably then the original consent is no longer valid. If this is the case, then new consent needs to be obtained*".

[30] For more examples of decisions which may, or may not, produce a legal or similarly significant effect on data subjects, see Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (6 February 2018), pp. 21-22, available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.

[31] Note that, where such substantial human intervention exists, the decision-making process can arguably be excluded from the scope of Art. 22 GDPR (as it is no longer fully automated). On this, see Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (6 February 2018, available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053), p. 30: "*The controller can still envisage a 'model' of decision-making based on profiling, by significantly increasing the level of human intervention so that the model is no longer a fully automated decision making process, although the processing could still present risks to individuals' fundamental rights and freedoms*", and p. 27: "*Human intervention is a key element. Any review must be carried out by someone who has the appropriate authority and capability to change the decision. The reviewer should undertake a thorough assessment of all the relevant data, including any additional information provided by the data subject*".

[32] See Section 2.1, above.

[33] For more on the applicability of Recital 33 GDPR to the use of consent in connection with scientific research purposes, see Article 29 Data Protection Working Party, *Guidelines on consent under Regulation 2016/679* (10 April 2018), pp. 28-30, available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051.

Inspired by this Recital, an innovative suggestion would be to **develop guidance and/or means for AI developers and users to provide dynamic information notices (using illustrations, flowcharts, videos, etc.) to data subjects, seeking to inform them about the key aspects of how their personal data will be used, walking them through the AI's process step-by-step and, where relevant, asking for their consent to the parts of the processing which are known at the time** – this information and consent request could then be updated/renewed in the case of any foreseen substantial changes at a later stage. However, in order for this to function in a manner similar to the possibility foreseen by Recital 33 GDPR, it is important that the renewal of consent is asked prior to the further processing which relies on it being carried out;[34] this would **require developers to design AI so that it does not automatically proceed with incompatible further processing of personal data**, unless it is confirmed – by the developer or user – that a legal basis for this exists.

Other issues arise specifically around the use of consent, such as the need to **allow for consent to be withdrawn**.[35] Developers must bear this in mind, and design AI-based systems to allow data pertaining to specific individuals to be extracted from a dataset and not further considered by the system in question. **Guidance and further research on how this can be attained in practice – in particular, considering that, where automated individual decision-making is concerned, Art. 22(2)(c) GDPR is, as our practical experience has shown, the most likely exception to be relied on – would be welcomed.**

[34] On this, note the position stated in Article 29 Data Protection Working Party, *Guidelines on consent under Regulation 2016/679* (10 April 2018, available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051), pp. 17-18: "*In any event, consent must always be obtained before the controller starts processing personal data for which consent is needed. WP29 has consistently held in previous opinions that consent should be given prior to the processing activity. Although the GDPR does not literally prescribe in Article 4(11) that consent must be given prior to the processing activity, this is clearly implied. The heading of Article 6(1) and the wording 'has given' in Article 6(1)(a) support this interpretation. It follows logically from Article 6 and Recital 40 that a valid lawful basis must be present before starting a data processing. Therefore, consent should be given prior to the processing activity.*"

[35] Art. 7(3) GDPR.

5 Challenges of Security

Security of datasets used in AI-based systems is a key concern.[36] There are several ways in which these **datasets can be maliciously compromised**, such as proprietary hacking of datasets, or even use of datasets against the AI in order to disrupt its decision-making.[37] Where machine-learning is concerned, the fact that such systems can autonomously deviate from their originally programmed goals can lead to the choices and predictions generated by such systems being misled by an attacker. The **impact of an integrity attack** on a dataset, or on an AI processing such a dataset, **can be massive, and could trigger public interest concerns** – consider, for example, where hacking a connected vehicle could put people's lives at risk. Security measures applied to AI must consider the direct risk that attacks on AI or its dataset may create for individuals.

In order to determine and implement appropriate security measures, AI developers and users must necessarily assess the relevant risks involved, so that they can select those measures deemed most adequate to address them. This refers to the risk-based approach promoted by the GDPR (in particular, for this case, Art. 32 GDPR), but which is also addressed in the NIS-D – as mentioned above,[38] the NIS-D expects OESs and DSPs (including those using AI) to manage the risks posed to their networks and information systems, through the implementation of appropriate security measures. If proper risk management is not carried out, then both the GDPR and NIS-D are breached. Once more, the manner in which it appears best to resolve this issue is **the development of further clear and understandable guidelines for AI developers and users on (1) AI risk management, and (2) examples of security measures,[39] at varying levels of sophistication (to caccount for developers and users of different sizes, types and economic capabilities), which may be considered in order to properly address identified risks.**

6 Related Cyberwatching.eu Publications

- Emerging technologies in the age of GDPR – Findings & recommendations from EU & R&I projects[40]
- Decentralized operation and security in the IoT Space[41]

[36] For more on this, see, e.g., Jake Saper, *How to* Hack *Your Way Into a Proprietary Data Set* (17 July 2018), available at: https://www.forbes.com/sites/insights-intelai/2018/07/17/how-to-hack-your-way-into-a-proprietary-data-set/.

[37] For more on this, see, e.g., Florian Tramèr et al, *Stealing Machine Learning Models via Prediction APIs* (August 2016), available at: https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf.

[38] See Section 2.2, above.

[39] Concerning use of AI and the NIS-D, one key reference to make is to the concept of SIEM (security information and event management), which indicates a model of approach to risk management combining two fundamental functions: (1) SIM (security information management) and (2) SEM (security event management). The key principle underlying any SIEM software solution is the ability to aggregate significant data from multiple sources, so as to identify deviations/anomalies from the norm, and then trigger appropriate actions to solve the security problem (e.g., when a potential critical event is identified, a SIEM solution can gather additional information, generate alarms and indicate additional security controls to block the progress of that event). By collecting and aggregating information from, e.g., servers, physical/virtual storage resources, PCs and smartphones, SIEM solutions essentially help to keep the various security measures which may be at a developer or user's disposal manageable. SIEM software can use heuristic algorithms that contemplate the probability of addressing cyber-attacks of various types, such as zero-day exploits, distributed denial of service (DDOS) attacks and brute force attacks. The system exploits a baseline, a basic model that allows it to perform pattern matching operations, log aggregation and analysis to locate anomalous activities. A solution of this importance can only be considered fundamental, in combination, in the most complex realities or more compliant with the requirements of the NIS-D, with the presence of a SOC (Security Operation Center).

[40] https://cyberwatching.eu/publications/emerging-technologies-age-gdpr-%E2%80%93-findings-recommendations-eu-ri-projects

[41] https://cyberwatching.eu/publications/decentralized-operation-and-security-iot-space

cyberwatching.eu has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 740129.