International Working Group
on Data Protection
in Telecommunications

675.57.14

Working Paper on Privacy and Artificial Intelligence[1]

64th Meeting, 29-30 November 2018, Queenstown (New Zealand)

Introduction

1. Artificial intelligence (AI) is high on the agenda of most sectors due to its perceived potential for radically improving services, commercial breakthroughs and financial gains. Over the next five years, it is expected that there will be mass implementation of AI across multiple sectors. However, enthusiasm for the opportunities offered by AI must be tempered by careful consideration of AI's impact on individual rights to privacy and data protection.

2. Recent advances in AI can be explained by the convergence of several factors including the development of innovative machine learning methods, the increase in available computational power and the availability of more labelled data, allowing the creation of complex statistical models.

3. AI systems in general, and machine learning technologies in particular, generally require the processing of huge volumes of data for their development. In a number of cases, this data is personal data, potentially impacting individuals' rights to data protection and to privacy.

Scope

4. The purpose of this working paper is to highlight the privacy challenges associated with the development and use of AI, and to provide a set of technical recommendations to help different stakeholders mitigate privacy risks when implementing it. While the use of AI also raises other ethical and societal concerns which deserve analysis, these are outside the scope of the present working paper.[2]

5. This paper focuses on some of the different ways in which AI interacts with personal data such as:
   - the use of personal data in algorithmic training/learning;
   - the application of AI to personal data (e.g., for decision-making purposes); and
   - the use of AI to extract personal data from data sets which superficially appear not to contain personal data.

[1] The Office of the Privacy Commissioner of Canada abstains from the adoption of this Working Paper.

[2] Internationally, Governments, Data Protection Agencies, and laws have variously sought to incorporate ethical frameworks, human rights controls and other guidance related to AI. While we encourage consultation with and implementation of appropriate guidance, questions of fairness and ethics are touched on but fall outside the scope of this paper.

6. This paper is intended for developers of AI systems, system providers, organisations purchasing and using AI systems, and for data protection authorities. [3]

Definitions[4]

7. Artificial Intelligence
AI is a term that has no universally accepted definition. AI can be described as the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

8. Specific vs. general AI
AI systems are currently developed for specific purposes. Some examples of the specific purposes for which AI is being used are: profiling, classification, image recognition, natural language processing and autonomous machines. So-called “general AI”, where one system is able to solve different types of problems, much like the human brain, is currently an unsolved challenge and this working paper will not address it.

9. Machine learning (ML)
ML is a subset of AI that uses statistical techniques to give computer systems the ability to "learn" from data with the goal of deriving an algorithm for the solution of a task without being given explicit instruction. The terms artificial intelligence and machine learning are often used as synonyms even though they are conceptually different.

10. Neural Network
Neural networks are largely inspired by our understanding of the way the human brain functions. These networks are built using what is basically a very simple component, an artificial neuron, which has a variable number of inputs and one output. Each input to an artificial neuron has a weight value that determines the extent of its influence on the final result. These values are adjusted when the network is trained to give the desired results.

11. Deep learning
Deep Learning is part of a broad family of machine learning methods based on data representations. As of today, it is almost exclusively based on neural networks. The deep part in deep learning is based on the number of layers found in the neural network. When a neural network has more than one (hidden) layer between the input and output layers, it is commonly viewed as deep.

How algorithms learn

12. There are three main forms of machine learning:
    - Supervised learning
    Supervised Supervised learning involves the use of labelled data. If the data includes images, the label may include information about the contents of each image, for example indicating if a dog or a cat is pictured.

[3] *Developers of AI systems* refers to private and public organisations and research institutions pursuing AI research and development. System providers are organisations and research institutions that use basic technologies developed by others (i.e., organisations that use AI in their own projects or in solutions supplied by others). These can be data controllers or merely a supplier of a service or product. Organisations purchasing and using AI systems may be both private and public organisations.

[4] This section is not intended to provide comprehensive or authoritative definitions of AI concepts, but seeks to delineate the understanding of these concepts that underlies the analysis and recommendations of this paper.

The data set is typically split in two, the larger part being used to train the model, the remaining part being used to test how precisely the model categorizes new data. The model requires a certain degree of generalisation to avoid overfitting. An overfitted model is too well adjusted, meaning it will perform very well with training data but poorly with new data.

Learning/training takes place as follows:

a. A set of labelled data is required.
b. Depending on data type, and what is considered relevant, the features (attributes of input data) to be used for learning are selected. The data is labelled to denote the correct prediction/answer.
c. A model is built that, based on the same features, will attempt to predict/produce the right label for unknown data.
d. The utility of the model is assessed using the part of the data set aside for that purpose. If results are unsatisfactory, the training process is renewed.

When in use after training, new and unlabelled data is fed into the system and a result is produced that should correspond with what the model learned during the training phase.

- Unsupervised learning

In unsupervised learning, the aim is to develop models that can detect patterns that would enable subsequent sets of unlabeled data to be clustered. If the training data consists of images of cats and dogs without any descriptive labels, the goal would be for these data to be sorted into two clusters sharing similar features – one consisting of images of dogs, and the other of cat images. However, the AI system will not be able to identify the nature of the two clusters, meaning that the system does not know that it sees images of cats and dogs.

Learning proceeds as follows:

a. A dataset is used in which there must be a certain number of similarities, or patterns, if it is to be meaningful.
b. The machine-learning algorithm will produce clusters based on similarities/patterns in the dataset.
c. A model is built that will sort, segregate, segment, cluster, etc. data using the patterns found during training.

When in use, the model will identify which group the new images belong to.

A disadvantage with this method is that the model cannot place data in groups other than those discovered during the learning process. It is therefore very important that the training data represents all possible clusters that new data might possibly belong to. Otherwise there is a risk that data could be forced into clusters where they do not belong, or that the data may simply not be clustered at all.

- Reinforcement learning

Reinforcement learning is based on trial and error. It allows machines and software agents to determine the optimal behaviour within a specific context in order to maximize performance - the model learns which actions are targeted towards the goal. While this reinforcement may use a pre-compiled set of data points as a starting point, the training phase may also proceed by immediately interacting with the real world domain, or acting upon computer generated responses. This means that less data, or no data at all, may be needed for the system to learn.

13. At some point during the learning process, there will be a need to assess the utility of the model to determine if it is meeting the specified requirements or goals (e.g., is the model consistently producing a correct/proper label for unknown data?). This can be done by measuring the quality of the segmentation, clustering etc. achieved by the model, by applying the model to new objects, and manually verifying the classification performed.

14. Different levels of complexity and intelligibility
There are several types of machine learning systems available or under development, each having different levels of complexity and intelligibility, which are adapted to solve various problems, and which require different amounts of data in order to learn.

15. A decision tree is one of the simplest and most popular forms of machine learning algorithms. Simply put, a decision tree is a tree in which each branch node represents a choice between a number of alternatives and each leaf node represents a decision. Decision trees train themselves, learning from given examples and predicting for unseen circumstances. The decision tree model might not be the best choice to analyse vast amounts of data, but it does offer a high degree of intelligibility. It is possible to follow the outline of the tree and see the criteria on which the result is based. With increasing amounts of data, however, a point will be reached where it will be difficult to obtain an overview and understanding of the decision-making process.

16. On the other end of the complexity scale, there are deep artificial neural networks. A neural network consists of a large number of artificial neurons arranged in more than two layers.[5] The models are based on weights and biases that are learned during training in something called backpropagation. Unlike normal programming statements or decision trees, the nature of the numeric values and the size of the networks can make it hard to understand how a decision is reached. When data is passed through the network, it is difficult to see how the information is combined and weighted to produce the final result.

17. Because some data must be viewed in context to make sense, for example words in the case of machine translation of speech transcription, some neural networks have a form of short-term memory. This allows them to produce different outputs based on the data that was processed previously, which of course makes it more difficult to determine how a result was derived. This also means that it can be very difficult to merely examine the algorithms to find out how they work and what decisions they reach.

Examples of AI in practice

18. Image recognition and analysis
Image recognition and analysis is an application of AI that has already been put to commercial use. These kinds of systems can recognize objects or people (e.g., image labelling or facial recognition), infer emotional states of people (e.g., facial gesture recognition) or detect and track a certain object or person through a video sequence.

[5] In 2016 Microsoft won an image recognition competition using a network consisting of 152 layers (https://blogs.microsoft.com/ai/2015/12/10/microsoft-researchers-win-imagenet-computer-vision-challenge)

Example:
An application which runs on Pivothead glasses aims to help visually-impaired individuals understand the world around them. When the wearer touches a sensor on the glasses, an image of their surroundings is captured, analyzed, and verbally described. For instance, if the system detects a person in the image, it can describe their approximate age, gender, facial emotion, and/or current activity. Similarly, if an image of text (a menu, news article, etc.) is captured, it is read back to the user. [6]

19. In facial recognition, a picture of a face is used to measure specific characteristics (i.e., nodal points on the face, such as the distance between the eyes or the shape of the cheekbones) and a template is produced. The trained algorithm compares this template to existing templates for categorisation, identification or authentication purposes. It should be noted that these systems are not infallible – they may incorrectly categorize, identify or authenticate an individual (i.e., a false positive error) or they may fail to categorize, identify or authenticate an individual (i.e., a false negative error).

Example:
The Chinese police have successfully tested smart glasses in conjunction with a facial recognition system to match travellers on a railway station with criminal suspects. According to the company that developed the technology, the system can identify faces from a database of 10,000 persons in 100 milliseconds. [7]

Example:
In 2015, it was revealed that the Google Photos service mistakenly tagged black people as “gorillas”. After the incident, the company promised “immediate action” to prevent any repetition of the error. That action has been to censor “gorilla”, as well as chimpanzee and monkey, from searches and image tags. That’s the conclusion drawn by Wired magazine, which tested more than 40,000 images of animals on the service. Photos accurately tagged images of pandas and poodles, but consistently returned no results for the great apes and monkeys – despite accurately finding baboons, gibbons and orangutans.[8]

[6] The Pivothead application is capable of more than describing an individual within the image or reading text. According to the case study linked via the Pivothead website the application can also describe scenery. http://www.pivothead.com/seeingai/

[7] The Independent, “Chinese police are using facial-recognition glasses to scan travelers”, 2018 http://www.independent.co.uk/news/world/asia/china-police-facial-recognition-sunglasses-security-smart-tech-travellers-criminals-a8206491.html

[8] Wired, “When it comes to gorillas, Google Photos remains blind”, 2018, https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Example:
In the UK, the police are experimenting with facial recognition technology. According to a report from Big Brother Watch, the police's use of this technology to recognize people is failing, with the wrong person being identified nine times out ten. South Wales police have been given two million pounds to test the technology, but so far it gets it wrong 91% of the time. The UK's Metropolitan Police used facial recognition at the 2017 Notting Hill carnival and the system was wrong 98% of the time, falsely telling officers on 102 occasions it had spotted a suspect.[9]

20. Natural language processing (NLP)
NLP systems use AI to allow people to interact with computers by speech or chat. This involves natural language and speech recognition and generation.

Example:
There are many products on the market using natural language processing. Some of the most popular are voice assistants like Google's Assistant, Apple's Siri, Amazon Alexa or Microsoft's Cortana, automated translation services like Google Translate, DeepL or Bing Translator or chatbots such as 1-800-Flowers and Swelly.

Note that models for speech recognition that are trained to recognize speech by individual speakers may contain personal data both on the semantic level (particular phrases used by a particular speaker), and on the phonetic level (manner of articulation of a particular speaker).

21. The development of AI systems capable of processing natural language makes it possible to collect and process data stored in audio or video recordings, or in images of text on paper. AI systems that interact with humans using natural language may use those interactions for learning and further development of the natural language capabilities of AI systems.

22. Autonomous machines
Autonomous machines are intelligent machines capable of performing tasks in the world by themselves, without explicit human control. Different features of AI may be applied in autonomous machines - natural language processing allows for direct interaction between humans and machines, while image recognition and audio analysis allow autonomous machines to recognize their environment. The accuracy requirements in decisions made by autonomous machines are often high.

Example:
Self-driving cars like the ones under development by Waymo, Uber or Tesla, home cleaning robots like Roomba or unmanned surveillance drones, are examples of autonomous machines.

23. Automated individual decision-making and profiling
AI systems and machine learning are increasingly used to automate individual decision-making and profiling. Profiling and automated decision-making can be very useful for organisations in many sectors, including healthcare, education, financial services and marketing. They can lead

[9] The Guardian, "UK police use of facial recognition technology a failure, says report", 2018, https://www.theguardian.com/uk-news/2018/may/15/uk-police-use-of-facial-recognition-technology-failure

to quicker and more consistent decisions, particularly in cases where a very large volume of data needs to be analysed and decisions made very quickly. Although these techniques can be useful, there are potential risks.

24. **Profiling** is any form of automated processing that uses personal data to evaluate certain aspects of an individual, such as personality, behavior, interests and habits to make predictions or decisions about them. Profiling may use AI and machine learning to create algorithms that find correlations between separate datasets, or between various personal attributes, and the observed behavior of individuals. These algorithms can be used to make a wide range of decisions, for example predicting behavior or controlling access to a service.

Example:
Several loan companies are using algorithms that factor in social media activity to determine whether to make a credit offer. A German company called Kreditech deploys a proprietary credit-scoring algorithm to process up to 20,000 data points on the loan applicant's social media networks, e-commerce behavior, and web analytics. Information about the applicant's social media friends is collected to assess the applicant's "decision-making quality" and creditworthiness.[10] In India and Russia, Fair Isaac Corp ("FICO") is partnering with startups like Lenddo to process large quantities of data from the applicant's mobile phone to conduct predictive credit-risk assessments. Lenddo collects longitudinal location data to verify the applicant's residence and work address, as well as analyzing the applicant's interpersonal communications and associations on social media to produce a credit score.[11]

25. **Automated decision-making** is the process of making a decision by automated means without any human involvement. These decisions can be based on factual data, as well as on digitally created profiles or inferred data. Examples of this include an online decision to award a loan or an aptitude test used for recruitment which uses pre-programmed algorithms and criteria. Automated decision-making often involves profiling. Not all profiling is used for decision-making purposes; nevertheless, the overlap between these two practices is considerable.

Example:
To predict the likelihood that a convicted person will reoffend, some correctional institutions in the United States use AI systems for profiling based on data on the convict's education, family background and social functioning, among other information.[12] Algorithms are also deployed in the criminal justice system to set bail, assess forensic evidence, and determine sentences and parole opportunities.[13] Several states use proprietary commercial algorithms, which may not be subject to open government laws, to make such determinations.

[10] https://www.kreditech.com/

[11] https://www.lenddo.com/

[12] See, for example COMPAS, https://www.cdcr.ca.gov/rehabilitation/docs/FS_COMPAS_Final_4-15-09.pdf or https://doc.wi.gov/Pages/AboutDOC/COMPAS.aspx

[13] See, for example, The New York Times, "Sent to prison by a Software program's Secret Algorithms", 2017, https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html, and EPIC.org, "EPIC – Algorithms in the Criminal Justice System", https://epic.org/algorithmic-transparency/crim-justice, for more general information.

Privacy challenges

26. The intensive use of data involved in many forms of AI, and the new data processing opportunities it brings, challenge fundamental data protection principles.[14] This paper highlights the most relevant challenges regarding privacy and the processing of personal data. While the use of AI also raises other ethical and societal concerns which deserve analysis, these are not within the scope of the present working paper.

27. Unlawful bias and discrimination
One of the major privacy challenges of AI systems is bias. Some data sets used to train machine learning-based and artificial intelligence systems have been found to contain inherent bias resulting in decisions that can unfairly discriminate against certain individuals or groups.

28. The fairness principle requires all processing of personal data to respect the data subject's legitimate interests, and that the data be used in accordance with what he or she might reasonably expect. The processing of personal data by an AI system may not respect the data subject's interests, or align with the data subject's reasonable expectations, especially if the algorithm is biased in some way, resulting in decisions or predictions with a discriminatory impact.

> **Example:**
> A research study found substantial disparities in the accuracy of three commercial face recognition systems conducting automated facial analysis. The study found that the commercial systems' training data were overwhelmingly composed of lighter-skinned subjects. The study showed that facial recognition of darker-skinned females had error rates of over 30 %, compared to an error rate of 0.8% for lighter-skinned males.[15]

29. How However, having a non-biased training dataset is not enough. For example, even if an AI system is not given input on sensitive attributes in an attempt to avoid discriminatory treatment, it is still possible for it to develop a compromised model on the basis of the information available that may in turn result in an unwanted discriminatory outcome. It is necessary to conduct an assessment of the results to make sure that there are no discriminatory effects.

30. There may be many reasons why individuals are underrepresented in data sets. For example, individuals may be very careful about what information they reveal about themselves, resulting in a lack of data to include in the data set. Similarly, they may not have access to or fluency in the technology that generates data about their activities and behaviors. Individuals may be deemed to be of less interest from a data perspective for some reason (e.g., they may not be in a certain economic class), resulting in their data not being included in the data set. Whatever the reason for their lack of inclusion, the result is that the AI system may exhibit bias against them.

31. AI systems use mathematically defined fairness and equity metrics to measure possible bias. The metrics used in the design of an AI system will deeply influence the outcomes of the AI system. However, it is not possible to design an AI system that is fair according to all metrics.

[14] See, for example GDPR Article 5.

[15] Buolamwin, Joy and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", Proceedings of Machine Learning Research 81:1–15, 2018, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Consequently, the metrics used should be part of the information provided to the users of AI systems in order to meet transparency obligations.

32. Even with concerted effort, it may be impossible to avoid bias in the output of an AI system. In some instances, there may not be any data available that is without inherent bias. Even when the most thorough efforts are made to avoid any bias in the selection of data, the data would reflect any bias which is present in social reality. In some fields, the state of knowledge may be insufficient to recognize bias in datasets or algorithms. Bias could also be ingrained in generally accepted principles of reasoning which have never been subject to thorough scrutiny. The lack of unbiased datasets in a domain should prevent the development of AI systems in that domain, as those systems will produce biased results.

33. Data Maximisation vs. the Principle of Data Minimisation
The ability to sift through and analyse vast amounts of data holds great potential for advancement in areas such as disease related research or personalised services across sectors. In the search for new connections and more precise analyses, it is tempting to give the system access to as much data as possible – this is sometimes called “data maximisation”. If the data used is personal data, this contradicts the principle of data minimisation.

34. The data minimisation principle requires that data be adequate, relevant, proportionate to the purpose for which it is collected, and limited to what is necessary for achieving that purpose. The capabilities that AI systems provide are pushing the limits for what is relevant, and the push to provide more and more data to facilitate connections pushes the data minimisation principle. Data may become newly meaningful in company with other data, greater processing capacity and deeper analyses. However, the potential for profit, research breakthroughs and more efficient services needs to be balanced with the potential risks and infringements that the extensive use of personal data has for individuals’ privacy and human rights.

35. Erosion of purpose limitation
The purpose limitation principle means that the reason for processing personal data must be clearly established and indicated at the time the data is collected. Furthermore, personal data cannot be re-used for incompatible purposes – uses must meet individuals’ reasonable expectations unless the re-use is explicitly mandated by law. This is essential if the individual is to have and exercise control over his/her information.

36. A challenge when developing AI is that it often requires many different types of personal data – information that in some cases has been collected for other purposes. Consider, for example, speech recordings made (on the basis of consent) for the improvement of the operation of voice-operated devices. These recordings could be used to train algorithms which seek to predict information about the health of the speaker. Such re-purposing of information may be useful and provide more accurate analyses than those which were technically feasible previously, but it can also be in contravention of the purpose limitation principle.

37. In addition, more powerful analytical tools might make it more tempting to use data for new purposes in order to enhance the output value. Economic and social benefits might be drivers to reuse data for new purposes.

38. Lack of transparency and intelligibility
Data protection is largely about safeguarding the right of individuals to decide how information about themselves is used. This requires that data controllers are open about the use of personal data, and provide necessary information about the processing.

39. Transparent artificial intelligence systems are ones in which it is possible to know how and why a system made a particular decision. The term transparency also addresses the concepts of intelligibility, and interpretability. Transparency enhances accountability.

40. It can be challenging to satisfy the transparency principle in AI powered decision-making systems. One reason is that the details behind an algorithm's functioning are often considered proprietary information, and so are closely guarded by their owners. Another reason is that, depending on the AI system, the algorithms might be so complex that even their creators do not know exactly how they work in practice.[16] This is AI's so-called black box problem.

41. If AI systems operate like black boxes and cannot be tested independently, the algorithms may be outside the scope of meaningful scrutiny and accountability. If organizations who use these systems for making automated decisions (as discussed earlier in the paper) are unwilling or unable to explain those decisions, then the individuals will have no way of knowing upon what information the decision was made, or whether the decisions were accurate, fair, or even about them. It will also be difficult for any individual to challenge or contest the decision. To protect individual rights in these situations, persons must be provided both the logic of the processing and an explanation of the automated decision-making.[17] Further, a lack of transparency and intelligibility in AI systems will also make it difficult for supervisory authorities (of any kind) to investigate, audit and inspect the systems.

42. Depending on the specific purpose of the AI, the inability to explain how a decision was reached could legally prevent the use of the system. For example, if social services denies an individual access to a social welfare benefit, many jurisdictions have a legal requirement to explain how that decision was made. The obligation to provide an explanation of how a decision is reached is also evident in some jurisdictions' privacy legislation.[18]

43. Finally, the black box problem makes it difficult to detect and remedy bias or security breaches in the processes. For example, detecting a data poisoning attack (poisoning the training data by injecting false data to compromise the learning process) may be impossible in a scenario where there is no explanation for the outcome of an AI system.

44. Erosion of consent
A lack of transparency in AI-powered systems, combined with a lack of intelligibility, may significantly erode the meaningfulness of consent. For consent to be valid, it shall be freely given, specific and informed. If individuals do not know how their data is going to be processed, and no-one can explain it to them, they will not be in a position to give a meaningful consent to the collection and processing of their data. In cases where a system does not depend on the availability of a specific individual's data, individuals could nevertheless face a loss of control over their data even if they refrain from providing consent. Examples of such situations include where individuals who have given their consent possess similar attributes as individuals who have not consented to the data processing, and yet this data is used to infer information or make decisions about those who have not consented.

[16] Creators know how their systems work theoretically (they implement methods such as gradient descent that should optimize the way the system work) but in practice the huge number of parameters and their automated tuning based on the statistical properties of the data make it hard to be able to precisely explain why such a decision was made, why such a parameter is so high while another is so low, etc.

[17] See GDPR Article 5 on Principles relating to processing (listing transparency as fundamental requirement for all processing of personal data); Article 13(2)(f) and 14(2)(g) on the Right to be informed; Article 15 on Right of access

[18] See for example GDPR Article 22 on Automated decision-making

45. Data analyses might uncover sensitive information
AI's vast analytical power is able to combine and analyse different information elements, which may not be sensitive in themselves, but when combined may generate a sensitive result. For instance, AI might identify patterns that can predict individual's dispositions, for example related to health, political viewpoints or sexual orientation. This kind of information is subject to special protection.

46. Similarly, inappropriate use of AI may weaken the effectiveness of consumer choice. By using data from consumers who opt in or decline to opt out of interacting with an AI system, that system can be employed to infer information about similarly-situated individuals who choose not to interact with those systems and do not share their data.

47. Risk of re-identification
Due to the ability of AI systems to process a wide variety of data from a multiplicity of sources, the use of AI may magnify the risk that individuals become identifiable in data sets, including data sets used for training purposes[19], which previously appeared to be anonymous. This is particularly true where data from different sources are combined and processed on a large scale. This makes anonymisation less likely to be successfully achieved. Thus it becomes ever more difficult to determine whether a data set is sufficiently and robustly anonymized, a process to which AI contributes for two reasons:

- The term "to identify" - and thus "to anonymize" - is complicated. Individuals may be identified in many different ways.[20] This includes direct identification, in which case a person will be explicitly identifiable by a single attribute (for example, their full name), and indirect identification, in which two or more data attributes describing physical, physiological, genetic, mental, economic, cultural or social characteristics must be combined in order to allow for the identification or singling out of individuals in a larger group. AI algorithms have the potential to uncover these characteristics from more idiosyncratic information (e.g. by uncovering a person's sexual identity from seemingly innocuous data); and

- Companies that use what is assumed to be an anonymized data set will not know for certain whether or not there are other external data sets available whose acquisition will make it possible to re-identify individuals in the data set. These acquisitions are becoming more routine in the quest for ever more powerful AI algorithms.

48. Information security risk
AI faces the risk of adversarial "injection" where third parties transmit malicious data to a learning AI system that in turns disrupts a neural network's functionality. For instance, a group of researchers confused an image recognition system by slightly modifying images used to train a system built to recognize road signs; the trained networks in question then misclassified almost all of the road signs a correctly trained algorithm recognized.[21]

49. The processing of personal data by an AI system in itself yields security risks. Like any other IT system, AI is vulnerable to security breaches. Searching out and exploiting those systems might be particularly attractive if malicious actors are able to access the large amounts of personal data that they contain, or the sources of such data.

[19] Veale M, Binns R, Edwards L., "Algorithms that remember: model inversion attacks and data protection law", Phil. Trans. R. Society, 2018, http://dx.doi.org/10.1098/rsta.2018.0083

[20] The Article 29 Data Protection Working Party, Opinion 05/2014 on "Anonymisation Techniques".

[21] Eykholt, Kevin et al., "Robust Physical-World Attacks on Deep Learning Models", Cornell University Library, 2017, https://arxiv.org/abs/1707.08945

50. Moreover, there are risks that emerge when AI systems can be reverse-engineered (i.e., when third parties “copy” the machine learning algorithm by replicating a model based on outputs or queries from the original system). In one example, after copying the algorithm, researchers were able to force it to generate examples of the potentially proprietary data from which it learned. If the algorithms are built on personal data, some of that information might become accessible as well.[22]

Recommendations

General Considerations

51. As declared by the International Conference of Data Protection and Privacy Commissioners[23], Artificial intelligence and machine learning technologies should be designed, developed and used in accordance with the principles of

    - fairness and respect of fundamental human rights,
    - accountability and vigilance,
    - transparency and intelligibility,
    - privacy by design and by default,
    - empowerment and respect of individual rights, and
    - non-discrimination and avoidance of biased decisions.

    All stakeholders, including researchers, developers and users of AI systems as well as legislators and regulators, should contribute to ensuring that the further evolution of AI systems is governed by these principles.

52. Fairness and respect of fundamental human rights
    Fairness and respect for human rights require that data are used only in in a manner consistent with the reasonable expectations of the individuals concerned, and only for purposes compatible with those for which they were collected, taking account of the impact not only on individuals, but also on groups and society as a whole, and ensuring that AI does not endanger human development. In short, there must be boundaries on uses of AI, and AI systems should not reflect unfair bias or make impermissible discriminatory decisions.

53. Accountability and vigilance
    Accountability and vigilance require the establishment of oversight mechanisms for audits, continuous monitoring and impact assessment of artificial intelligence systems, and their periodic review, collective and joint responsibility of all actors and stakeholders, awareness raising, education, research and training, and where necessary the involvement of trusted third parties or independent ethics committees.

54. Organizations, not algorithms, are accountable for the results of all data processing involving the use of AI-based systems or services. To help ensure accountability, roles and responsibilities must be clearly defined, assigned and well documented.

55. In instances where an organization is using an AI-based service provided by a third party, the respective roles, responsibilities and rights of the organization and supplier with respect to the

[22] Wired, “How to steal an AI”, 2016, https://www.wired.com/2016/09/how-to-steal-an-ai/

[23] 40th ICDPPC – Brussels, 2018, Declaration on Ethics and Data Protection in Artificial Intelligence, https://icdppc.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf; see also Universal Guidelines for Artificial Intelligence (2018), https://thepublicvoice.org/ai-universal-guidelines/

processing of personal data, including those related to the security of the AI systems, should be clearly articulated and allocated.

56. Organizations need to demonstrate that they are being accountable and can make responsible and ethical decisions regarding their use of AI-based services. Both models and their underlying algorithms require continuous assessment. This necessitates regular audits to ensure that decisions resulting from the profiling are responsible, fair, ethical and compatible with the purpose(s) for which the information was collected and is being used.

57. Transparency and intelligibility
Transparency and intelligibility require scientific research on explainable artificial intelligence, the development of innovative ways of communicating relevant information, transparent practices of organizations and of algorithms, auditability of systems, appropriate information to individuals so that they are aware when they interact with AI systems or provide data to them and overarching human control of the systems.

58. AI-based systems and services should be designed to support internal and/or external audit or review. As far as possible, AI-based systems and services should be based on data, algorithms, models, protocols, designs and implementations that are open for external review and/or testing. Open audits, or audits by trusted entities, can help to provide assurance that the AI-based services do in fact have all the claimed properties and will not generate unfair or discriminatory outcomes.

59. Where possible, AI-based systems and services should be based on data, algorithms, models, protocols, designs and implementations that are as intelligible as possible. A number of promising techniques have been proposed including:

- Explainable AI (XAI)[24]: XAI is the idea that all the automated decisions made should be explicable. With people involved in a process, it is often desirable that an explanation is given for the outcome. As an example, there is a project underway in this field, being run by the Defense Advanced Research Projects Agency (DARPA), where the objective is to gain more knowledge about providing understandable explanations for automated decisions;

- Local Interpretable Model-Agnostic Explanations (LIME)[25]: LIME is a solution that produces explanations ordinary people can understand. In the case of image recognition, for example, it will be able to show which parts of the picture are relevant for what it thinks the image is. This makes it easy for anyone to comprehend the basis for a decision; or

- Counterfactual explanations[26]: These are explanations that describe the smallest change to a variable used by the algorithm (like income, test scores, or account activity) that would be needed for the algorithm to arrive at a desirable outcome. As multiple variables or sets of variables can lead to one or more desirable outcomes, multiple counterfactual explanations can be provided, corresponding to different choices of nearby possible worlds for which the counterfactual holds. Counterfactuals describe a dependency on the external facts that lead to that decision without the need to convey the internal state or logic of an algorithm. These explanations thus aim at informing and helping the individual understand why a particular

[24] DARPA, “Explainable Artificial Intelligence (XAI)”, https://www.darpa.mil/program/explainable-artificial-intelligence

[25] Tulio Ribeiro, Marco, Singh, Sameer, Guestrin, Carlos, “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”, Cornell University, 2016, https://arxiv.org/abs/1602.04938

[26] Wachter, Sandra and Mittelstadt, Brent and Russell, Chris, “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR”, Harvard Journal of Law & Technology, 31 (2), 2018. SSRN: https://ssrn.com/abstract=3063289 or http://dx.doi.org/10.2139/ssrn.3063289

decision was reached, providing grounds to contest the decision if the outcome is undesired, and understanding what would need to change in order to receive a desired result in the future.

60. In relation to AI-based systems or services, information regarding the categories of information collected, the purposes for which the information will be used, the identity of the actors involved in the processing, how long the data will be retained and the general security practices that are in place, should be published. This information should be kept up-to-date and should be clearly communicated to relevant individuals.

61. Privacy and Ethics by Design
Privacy and ethics by design and by default require implementing technical and organizational measures and procedures, assessing and documenting the expected impacts on individuals and society and identifying specific requirements for ethical and fair use of the systems and for respecting human rights as part of the development and operations of any artificial intelligence system.

62. AI-based systems and services should be developed and designed in accordance with privacy and ethics by design principles.

63. AI-based systems and services should be subject to an independent ethics review mechanism, either internal or external[27], to ensure that the proposed AI system or service will behave in an ethical manner.

64. AI-based systems and services should be subject to extensive testing to ensure that any regulatory or ethics-related design issues related to the product or service are identified and addressed in a timely manner.

65. AI-based systems and services should be subject to a privacy impact assessment and a risk analysis at appropriate stages of their lifecycle (e.g., development, implementation, decommissioning). The necessary technical and organizational measures, identified during these analyses, should be implemented.

66. Empowerment
The opportunities offered by AI should be used to foster equal empowerment and enhance public engagement. This means respecting and facilitating the exercise of individuals' rights to data protection and privacy such as the rights to access to information, to object or request erasure of information, as well as the rights of freedom of expression and information, non-discrimination and, where applicable, individuals' right not to be subject to a decision based solely on automated processing or the individuals' right to challenge such decision.

67. Those who are subject to an automated decision by an AI-based system or service should be informed that they have been subject to such a decision and should have the opportunity to fully understand the reasoning behind that decision as well as the factors that (most) influenced the decision. Any associated automated decision-making or other rule-based systems and the reasoning underlying determinations made with or by those systems must be explainable to individuals and organizational users in clear, simple, and easy to understand language.

68. Non-discrimination
Non-discrimination and the avoidance of biased decisions require recognition of and respect for

[27] See, for example, Trilateral Research, "Research ethics for industry 4.0", https://trilateralresearch.co.uk/research-ethics-for-industry-4-0/, or O'Neil Risk Consulting and Algorithmic Auditing (ORCAA), at http://www.oneilrisk.com/ for examples of a commercially available "ethics board".

international legal instruments on human rights, research into technical ways to identify, address and mitigate biases, ensuring that the personal data is accurate, up-to-date and as complete as possible, and specific guidance and principles in addressing biases and discrimination.

69. There may be several forms or stages of training for AI systems (e.g., initial training during development, acceptance testing during implementation, and ongoing training during use). At all stages of an AI-based system and service's lifecycle, steps should be taken to ensure that training data is of the highest quality and relevance possible. This includes ensuring the data is as correct, accurate, complete, relevant, representative and up-to-date as possible. It also includes ensuring that, to the greatest extent possible, the data is free from bias based on race, age, gender, sexual orientation, religious belief, income level, or other protected grounds.

Specific Considerations

<u>Developers (of AI components, systems and services)</u>

70. Developers should ensure that the purposes for which they are processing personal data (e.g., system training to enhance face recognition) are clearly defined, well documented and correspond with the expectations of individuals about the use of their information.

71. Developers should minimize, to the greatest extent possible, the amount of personal data used during development, and ensure that any such data is limited to that which is relevant and necessary for the defined purposes (e.g., training). In addition to techniques such as the use of anonymized data, several possible techniques have been identified that may enable this minimisation including, but not limited to:

- Generative Adversarial Networks (GANs)[28]: GANs are used for generating synthetic data. As of today, GANs have mainly been used for the generation of images. However, it also has the potential for becoming a method for generating huge volumes of high quality, synthetic training data in other areas. This may satisfy the need for both labelled data and large volumes of data, without the need to utilise great amounts of data containing real personal information;

- Federated Machine Learning[29]: This is a form of distributed learning. Federated learning works by downloading the latest version of a centralized model to a client unit, for example a mobile phone. The model is then improved locally on the client unit, on the basis of local data. The changes to the model are then sent back to the server where they are consolidated with the change information from models on other clients. An average of the changed information is then used to improve the centralized model. The new, improved centralized model may now be downloaded by all the clients. This provides an opportunity to improve an existing model, on the basis of a large number of users, without having to share the users' data;

- Transfer Learning[30]: it is not always necessary to develop models from scratch. Existing models that solve similar tasks can be utilized. By basing processing on these existing

[28] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, "Generative Adversarial Nets", Département d'informatique et de recherche opérationnelle Université de Montréal, https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[29] Google AI Blog, Federated Learning: Collaborative Machine Learning without Centralized Training Data, 2017, https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[30] Machine Learning Research Group, University of Texas at Austin, http://www.cs.utexas.edu/~ml/publications/area/125/transfer_learning

models, it will often be possible to achieve the same result with less data and in a shorter time. There are libraries containing pre-trained models that can be used; and

- Matrix capsules[31]: Matrix capsules are a new variant of neural networks, and require less data for learning than the current norm for deep learning. This is very advantageous because a lot less data is required for machine learning.

72. Where possible, developers should make the training data available for external review and/or testing. Where this is not possible (for instance, if the data is business sensitive), organizations should be able to clearly demonstrate that they have taken steps to ensure the quality and relevance of the data, either internally or through the use of a reputable third party.

<u>Providers (of AI components, systems and services)</u>

73. Providers should ensure their systems include mechanisms and techniques that will support compliance with relevant privacy regulation and document how these requirements are met. Documentation is one of the requirements of the regulations, and may be requested by customers, users or oversight bodies.

74. As appropriate, providers should ensure that data used for marketing and sales purposes, or as part of acceptance testing, is as correct, accurate, relevant, representative, complete and up-to-date as possible.

75. Providers should ensure that their algorithms, data, protocols, designs and implementations are open for external review and/or testing. Open audits, or audits by trusted entities, can help to provide assurance that the AI-based systems or services in fact have all the claimed properties and will not generate unfair or discriminatory outcomes.

<u>Organizations (implementing and using AI systems or services)</u>

76. Organizations intending to use AI-based systems or services should ensure that they have an appropriate legal basis for the processing of personal data.

77. Organizations intending to acquire or use AI-based systems or services should specify their privacy and data protection requirements, as well as any additional requirements (e.g., with respect to transparency, and auditability). These requirements should be clearly documented (e.g., in a contract with an AI developer).   Organizations should make these requirements known to providers and developers as appropriate.

78. Organizations should only engage providers of AI-based systems or services that offer sufficient guarantees that the privacy and data protection rights of individuals are adequately protected and that other requirements are adequately addressed.

79. Organizations should ensure that any data used for ongoing training, testing or evaluation of an AI system or service is as correct, accurate, relevant, representative, complete and up-to-date as possible.

80. Organizations should not collect, use, or disclose personal information in ways that would run counter to the context in which individuals provided that data. Organizations wanting to use the collected data for a purpose different than the original one must assess the compatibility between the original and the new purposes on a case-by-case basis.

[31] Hinton, Geoffrey, Sara Sabour and Nicholas Frosst, “Matrix capsules with em routing”, Google Brain, Toronto, Canada, 2018, https://openreview.net/pdf?id=HJWLfGWRb

81. Organizations should only process as much personal data as they need to complete specified purposes. Organizations should minimize the amount of personal data used by the system or service. This minimisation may be achieved through a number of techniques (e.g., removal of the personal data from the data set, using synthetic data instead of actual data, anonymization and so on).

82. Organizations should ensure appropriate transparency regarding the use of algorithms and profiles that may influence decision-making. Any associated automated decision-making or other rule-based systems and the reasoning underlying the determinations made with those systems must be explained to individuals and organizational users in a clear, simple, easy to understand and timely manner.

83. In cases of automated individual decisions, individuals should know the decision was automated, and have access to the decision and its reasoning in order to assess whether their information has been processed fairly. Organizations should implement innovative, practical and expedient procedures that facilitate a human evaluation of decisions in cases where a different point of view is submitted, counter-arguments are presented, or where the decisions are challenged.

Data Protection Authorities

84. Data Protection Authorities (DPAs) should ensure that they possess sufficient knowledge and expertise in order to give guidance and to investigate possible breaches of relevant data protection or privacy regulation. This may be achieved through the acquisition of expertise by DPA staff or by ensuring access to relevant external expertise through partnerships with academia, industry, NGOs and other government agencies as appropriate.

85. DPAs should strengthen their awareness raising activities by providing guidance to relevant stakeholders. This could include promoting the application of privacy by design principles with AI-based services developers, providers and users.

86. DPAs should support the implementation of codes of conduct, data protection and privacy certification schemes, as well as the development of suitable data protection and privacy impact assessment frameworks and tools, in order to foster the development of privacy friendly AI-based systems and services.

87. DPAs should also strengthen their supervisory activities. This could include supporting the development of international arrangements for enforcement cooperation and the conduct of joint enforcement activities. It could also include auditing the development, implementation and use of AI systems and services in order to identify practices that create risks for individuals. As appropriate, DPAs should share the results of these audits with other regulatory authorities.