Draft for Consultation

UN Special Rapporteur on the Right to Privacy

Data Privacy Guidelines for the development and operation of Artificial Intelligence solutions

in accord with the UN Charter on Human Rights and International Covenant on Civil and Political Rights

A. Background

1. Purpose

1.1. The purpose of this paper is to provide guiding principles concerning the use of personal and personal related information in the context of Artificial Intelligence (AI) solutions[1] developed as part of applied Information & Communication Technologies (ICTs), and to emphasise the importance of a legitimate basis for AI data processing by governments and corporations.

1.2. This Guidance is intended to serve as a common international minimum baseline for data protection standards regarding AI solutions, especially those to be implemented at the domestic level, and to be a reference point for the ongoing debate on how the right to privacy can be protected in the context of AI solutions.

1.3. AI solutions are intended to guide or make decisions that affect all our lives. Therefore, AI solutions are currently subject to broader debates within society. The subject of these debates - moral, ethical and societal questions including non-discrimination and free participation, are still to be solved. All of these questions are preconditioned by lawful data processing from a data

[1] There are several definitions of Artificial Intelligence. The meaning intended here is the most common one and exemplified by that of the Oxford dictionary which defines Artificial Intelligence (AI) 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.“ This is far from being an exhaustive list of applications of AI technologies.

privacy perspective. The data privacy underpinnings for AI solutions are the focus of this Guidance.

1.4. This guideline is based on the United Nations Charter of Human Rights (The Universal Declaration of Human Rights, Dec. 10th, 1948, reaffirmed 2015, UDHR) and reflects the spirit as well as the understanding of this Charter. Above all Article 7 (non-discrimination) and Article 12 (right to privacy) shall be considered whenever developing or operating AI solutions. The themes and values of these UDHR Articles are found in Articles 2 and 3 (non-discrimination), and Article 17 (privacy) of International Covenant on Civil and Political Rights, and are obligations upon countries that have ratified the Treaty.

2. Scope

2.1. This Guidance is applicable to the data processing of AI solutions in all sectors of society including the public and private sectors. Data processing in this context means the design, the development, the operation and de-commissioning of an AI solution.

2.2. This Guidance is applicable to all controllers of AI solutions. Controller in this context means designer, developer or operator (self-responsible or principal) each in its specific function.

2.3. This Guidance does not limit or otherwise affect any law that grants data subjects more, wider or in whatsoever way better rights, protection, and/or remedies. This Guidance does not limit or otherwise affect any law that imposes obligations on controllers and processors where that law imposes higher, wider or more rigorous obligations regarding data privacy aspects.

2.4 This Guidance does not apply to AI solutions that might be performed by individuals in the context of purely private, non-corporate or household activities.

Draft for Consultation

Artificial Intelligence and Data Privacy

1. Introduction

Current AI systems represent a combination of analysis systems based on formalised expert knowledge (Data Warehouse, Business Intelligence) and machine learning as well as the targeted application of what has been learned. There is a differentiation between pre-programmed, algorithmic systems for the solution of specific problems, and systems that can learn. The latter are equipped with learning algorithms and have to be trained.

In the algorithmic decision-making process, which is regularly used as the basis for AI, an assessment is made based on information, which leads to a decision, forecast or recommendation for action. In the case of “supervised learning” the AI system has solution criteria for solving a specific problem, whereas in case of a “non-supervised learning” the AI-system itself will recognise the relevant solution criteria.

Consequently, the data processing as well as the decision made as a result of this processing, have potential risks for the data subject.

The classical IT with its elements "input" - "processing" - "output" is extended by the abilities "perceiving" - "understanding" - "acting" - "learning". These characteristics, which until now have only been undertaken by humans, can now be performed also by machines to an increasing extent. The term "understanding" is new territory in connection with computers and must be accompanied by critical review of traceability and adherence to ethical values.

Machine learning refers to a series of optimisation methods in artificial neural networks, among others. AI systems can have very complex structures between the input and output layers. By mapping several hierarchical processing layers, machine learning can become considerably more efficient (Deep Learning). However, this inevitably results in a loss of traceability in AI decisions. Due to the complexity of the algorithms and the multitude of arithmetic operations performed by the machine, the deeper processing layers (hidden layers) elude transparency in the decision criteria and their weighting.

Although the disclosure of the algorithms on which the AI is based is a core demand in the current debate about more transparency in AI systems, the concrete verification of the decision logic of highly complex AI systems on the basis of disclosed algorithms is likely to be difficult in practice. "Explainable AI systems" is an approach that is currently being intensively researched. To all extent possible and practical, the users of AI systems need to disclose the

decision criteria and their weighting for decisions, beside the factual results of an AI based data processing. The purpose, overall functions, supporting processes, data sources used to enable the range of the outcomes need to be documented and explainable (“Explainability”) to be able to manage the inherent risks. In the case of a failure in process or outcomes, it would be possible therefore, to capture digital evidence to be able to reconstruct what happened and why.

Monitoring the decision-making processes of AI systems from "outside", by reviewing the decisions themselves against a pre-determined purpose of the system and ethics governance has many benefits, including practicality.

AI decisions falling outside the expected range of outcomes or decisions can be identified and an intervention made. Tools developed specifically for the detection of unexpected outcomes and for analysis of AI decisions are one pre-requisite. Monitoring machines exclusively by machines increases the possibility of unforeseen risks or “unknown unknowns”. This necessitates the principle that human judgements must always dominate AI monitoring processes.

In addition to the efficiency of the learning mechanisms, successful machine learning depends on the quantity and quality of the available data. The "Big Data" trend in IT and the increasing mass availability of data of a high quality are currently significantly accelerating the development of AI systems.

Transparency of the data sources used and the lawfulness of their processing in AI systems are therefore key data privacy requirements.

The very complex psychological and emotional processes of human knowledge and decisions are likely to remain the domain of humans, rather than machines, for some time to come. Therefore, when evaluating and weighing up data privacy law in relation to AI systems and their decision making, it must be borne in mind that machine decisions are based on different principles and mechanisms (although developed largely by humans) than those applied to human decisions.

In order to achieve the necessary security for AI systems, comprehensive ethical and legal governance for AI decisions must be effectively implemented in the control environment of an entity making use of AI solutions.

2. Data Privacy Principles for the use of AI solutions

Irrespective of the jurisdiction or the legal environment applying to the controller, seven main principles are mandatory considerations in the planning and implementation of AI solutions:

2.1 Jurisdiction

2.2 Lawful basis and purpose limitation

2.3 Accountability

2.4 Control

2.5 Transparency and 'Explainability'

2.6 Rights of the "Data Subject"

2.7 Safeguards

The following specification of the seven principles do not replace any other or stricter data privacy regulation applicable to the controller working with an AI solution. General data privacy requirements may apply also. There may also be additional requirements for the processing of specific kind of data such as health data.

3. Considering Ethical Aspects

There is a responsibility upon our society to develop AI solutions ethically and responsibly. AI solutions now, and increasingly will affect many areas of our daily lives and will have a deep influence on our personal living and working situations. AI solutions will have to cover a broader range of fundamental principles reflecting both legal and ethical questions. What is important is what we *do* with this technology.[2]

Non-discrimination is critical to avoiding inequality, injustice and suffering. It needs to be accurately monitored and any occurrences corrected to avoid adverse effects. Human rights assessments should always be undertaken alongside data privacy assessments to provide a holistic overview of the necessary framing conditions. Several Committees around the world are currently working on drafting Codes of Ethics for AI solutions. Reference should be made to them.

[2] Walsh, T. (2018) *2062 The World that AI Made,* La Trobe University Press/Black Inc.

B. The Data Privacy Principles for AI solutions

1. Jurisdiction

To create legal certainty and traceability, AI solutions should be implemented and operated in a single jurisdiction, and that jurisdiction should have suitable legislation for best practice governance and risk management of AI. Where an AI solution uses a distributed decision-making mechanism, that distributed mechanism should also be in a single jurisdiction.

Unless and until there is developed a specific ad hoc international law mechanism for settling jurisdictional issues in ICT and especially AI solutions developed in one jurisdiction but used in another, where an AI solution is required to operate across multiple jurisdictions, it should be implemented and operate as a multinational federation of individual single jurisdiction AI solutions.

2. Lawful basis and purpose limitation

As the processing of personal data of individuals always intrudes into the rights of the data subject, an AI solution must have a sound legal basis if it deals with personal data. This becomes even more important if the processing itself is designed to lead to, or to make decisions affecting the position or the rights of the data subject. Irrespective of the jurisdiction or the controller's individual legal environment, one or more of the following case categories may be utilised as a sufficient legal basis for the processing of personal data by an AI system:

2.1 Specific legal basis at law if the law was drafted in accordance with democratic principles and generally accepted human rights, and if the law is addressing the conflict of interests between controllers and the data subjects.

2.2 If the usage of the AI solution is needed for the fulfillment of a contract with the data subject and if this contract does not disadvantage the data subject materially.

2.3 If the data subject has given the free, uninfluenced consent on an informed basis. The consent has to be given by concrete action and the controller must provide a consent management system that allows withdrawal of the consent at any time and includes adequate documentation.

2.4 On the basis of a legitimate, overweighing interest of the controller if the data subjects are adequately informed before the processing starts and are

given the opportunity to object to the processing within a reasonable time period.

2.5 Every AI solution is bound by and limited to the purpose for which it was originally designed and correctly documented. Other or additional uses (such as further processing) or the usage by another controller need to be evaluated anew with regard to their legal basis and safeguarding measures.

3. Accountability

Each AI solution needs either a legal or a natural person that takes the full responsibility for the data processing and its results. This covers all aspects of the management of the process and the technology including the lawfulness of the processing, the documentation of the processing and the adaption of the processing, the results of the processing and the fulfillment of the rights of the data subjects.

These responsibilities, including an eventual processor of the AI solution if not identical with the controller, must be transparent and adequately accessible by the data subjects as well as for public supervisory authorities and regulators.

Appropriate governance, particularly in larger legal entities, requires the establishment of a Data Privacy Officer. The functions of this role include responsibility for advice on compliance with data privacy requirements and for monitoring the implementation of the AI solution. The role must be provided with adequate resources and authority to undertake these functions.

4. Control

AI solutions must be under full control of the controller. From the first design idea until the final switch off and decommissioning, it must be clear what data are processed in the AI solution, what parameters and data quality metrics provide the basis for the decision making and how they shall be balanced and weighted against each other. The results must be monitored continuously and corrected if necessary. In the area of automated decision-making solutions, no decisions are to be made based on conscious or unconscious bias. Possible bias and discriminating effects must be checked and corrected both before roll-out of a system and at regular intervals throughout its lifetime.

In the case of AI for decision support systems, a similar set of controls is required to avoid incorrect proposals for the decision maker.

The controller must be able to stop or change the processing at any time. Incorrect results and the corrective measures must be documented as well as be taken, to mitigate any risks for the data subjects. Once their use for

identification, corrective or forensic purposes, the false results must be deleted without undue delay.

A communication channel that allows employees and people from outside the entity to address any kind of critical findings regards the AI solution or its results, is necessary.

5. Transparency and Explainability

AI solutions must be made transparent to the public and the data subjects. The information must cover all relevant aspects that might be of interest regarding the evaluation of the solution and possible rights of the data subjects. This includes “Explainability” of the purpose, the overall functions, supporting processes, used data sources and the range of the planned outcome. These may include *inter alia*:

4.1 Data sources and data used to feed the AI solution as well as data resulting from the AI solution.
4.2 Purpose and legal basis for the processing.
4.3 Parameters that build the basis for AI decisions and their weighting.
4.4 Clarification on whether the AI solution is intended to prepare decisions to be finally made by human beings (decision support) or if it is making the final decision itself (automated decision making).
4.5 Responsibilities within the controller and processor – if not identical with the controller - and contact details and possible communication channels
4.6 Integration of third parties (e.g. other controllers or processors) and transfer to other countries (if so) as well as the reason for the integration and the transfer. This also requires a declaration that third parties are bound by the same requirements as the controller no matter where in the world they are allocated.

The necessary information must be published in the data privacy policy referring to the AI solution and must be accessible and understandable in the way that is relevant to the data subjects.

6. Rights of the “data subject”

Persons whose personal information or identifiable personal information are processed by the AI solution (data subjects) shall have the following rights:

5.1 Right to withdraw consent without negative consequences if consent was given and utilised as the legal basis for processing.

5.2 Right to object to the data processing for good reason at any time if the processing is based on “legitimate interest” (Section C, 2.4).
5.3 Right to information regarding the fulfillment of all Data Privacy requirements listed in this Section.
5.4 Right to proportionate access to their data with comprehensive written information about their personal data and how their personal data is used and processed as well as the results and the way the results might affect the position and individual rights of the data subject.
5.5 Right to request a decision by a human being if they have reasonable doubts that the decision proposed or made by the AI solution is not accurate or correct.
5.6 Right to correct data if it is incorrect.
5.7 Right to make a complaint if they have a good reason for that.
5.8 Right to erasure and purge the data if the purpose of the AI solution ceases to exist or if the data is no longer needed for another legal purpose.

The rights listed in this paragraph do not overrule other rights and/or exceed rights granted to the data subjects under the applicable law in a given jurisdiction.

7. Safeguards

AI solutions shall function in a robust way and shall be secured by appropriate safeguards against risk, using methods that foster trust and understanding across all parties involved, including the data subjects and the public. This means that all AI solutions must, before deployment, even if only on a test basis, undergo at a minimum, a data protection risk assessment that identifies the specific risks and criticalities associated with the intended solution.

Using a “privacy by design” approach, technical and organisational safeguards to mitigate the identified risks must be assessed individually; this should cover measures like anonymisation or pseudonymisation, encryption, client separation, access management (limitation), deletion policy, log and activity monitoring.

Emerging new risks and challenges arising from technological, architectural and/or structural developments, like distributed computing, must be considered and assessed during the risk assessment.

The risk mitigation additionally can be based on international standards such as ISO 27000 series (information security management systems), particularly ISO 27701, which contains data privacy extensions. It must contain at the minimum:

- Protection measures: Controls and measures to protect against the effects of assessed risks.

- Detection measures: Controls and measures to detect abnormalities as soon as possible.
- Responding measures: Controls and measures to contain and defeat the risk or abnormal event, and to ensure that core business processes can still function, until such time as the overall solution recovers to normality.

C. Assessment of Criticality of AI Solutions

The measures to be taken must be proportionate to the risks associated with infringements of human rights, especially non-discrimination, as well as the complexity or criticality of a data processing solution.

There are several suitable approaches and this Guidance provides examples of risk handling.

1. Human rights Assessment in the planning phase

All AI solutions must respect the rule of law, human rights, democratic values and diversity. Therefore, every planned AI solution shall undergo a timely human rights assessment. In particular, the rights of equal treatment and equal share shall not be unlawfully violated by the planned AI solution.

One possible scenario might be that AI solutions are using information that is a result of an unconscious bias and therefore will lead to results that might discriminate against certain people or parts of our society. It can also be that an AI solution, fed with the 'right' information will lead to 'wrong' results as the learnings of the AI solution derived from the collected information might lead to erroneous assumptions by the AI solution.

Whereas privacy by design and by default and a Privacy Impact assessment should be adopted as basic requirements for any AI-based system, the risks outlined above and others mean that an essential part of the AI planning is to carry out a prior assessment of how any human rights and not only privacy might be affected by the implementation of the AI solution.

2. Test and Correction phase – monitoring

After the planning phase and the initial high-level human rights assessment, the identified framing conditions must be considered in the further development phase. During the implementation phase and before going live, AI solutions should undergo an intensive "test run" with testing data in a separate, self-contained environment to assess if the underlying general

assumptions are not only considered but fulfilled. Only if the controller can be sure that the AI solution runs properly, should it be launched for live operations.

During the whole runtime of the AI solution (until the final “switch off”) the results produced by the AI solution must be monitored against the fundamental requirements defined in the planning phase.

The difficulties of controlling all aspects of the algorithms’ operations and the constant change of algorithms during the runtime of an AI solution, make it essential that the results are constantly checked against the initial intended purpose of the AI solution in another feasible way to provide a point of comparison. If a deviation is suspected or observed, the data feed for the AI solution must be adapted accordingly or the solution itself must be stopped.

To gain the benefits of new creative approaches, and to widen the horizon of the developer and the controller, input and feedback from the ‘data privacy’, civil society and user communities needs to be factored into the development, testing and monitoring of AI solutions. For testing purposes, ready to run AI solutions can be provided by installing a so-called black box in the internet. In this scenario, the separated and self-contained solution is open to third parties to input data to ascertain the type of results the AI solution will produce.

3. Criticality assessment based on the usage of different kind of data

Besides proper planning, testing and implementation, the criticality of data and the intended purpose are relevant also to the measures necessary for proper processing.

This applies to general data, like general personal information or data in the context of telecommunication services or data in a health context. Health data and some other information, for example, the contents of telecommunications, have to be treated more rigorously than less sensitive personal information. This means that the relevant technical and organisational measures must be stronger and more rigorous than in other cases (for example, strict purpose limitation and data minimisation, encryption, pseudonymisation, restricted access and early deletion or anonymisation).

The context of the intended data use plays a key role in determining the level of protection required. If the controller uses general personal information purely for storing purposes this might be less critical than using it for profiling

purposes and/or marketing. Accordingly, the legitimacy of the purpose and the safeguarding measures must be assessed very carefully.

These actions must be undertaken and documented during the data privacy risk assessment.

D. Additional considerations

1. External audits and certification
   External certification of an approved auditor in data privacy who is also formally recognised as having AI expertise should be considered. This may be helpful in allaying the concerns of the public and those of the data subjects. This may be particularly applicable for AI solutions that could lead to major adverse outcomes and a loss of trust by the public and/or the regulatory community.

2. Around the world, new legislation and regulations are being considered that will affect the majority of AI solutions. Compliance with these will largely depend on:

   a. Compliance with existing and emerging national and international standards
   b. Certification by an appropriate certification authority operating under a national or international agreement.

3. Those responsible for AI strategies and/or or operational AI solutions should closely follow the variety of discussions occurring about AI and associated ethical questions.