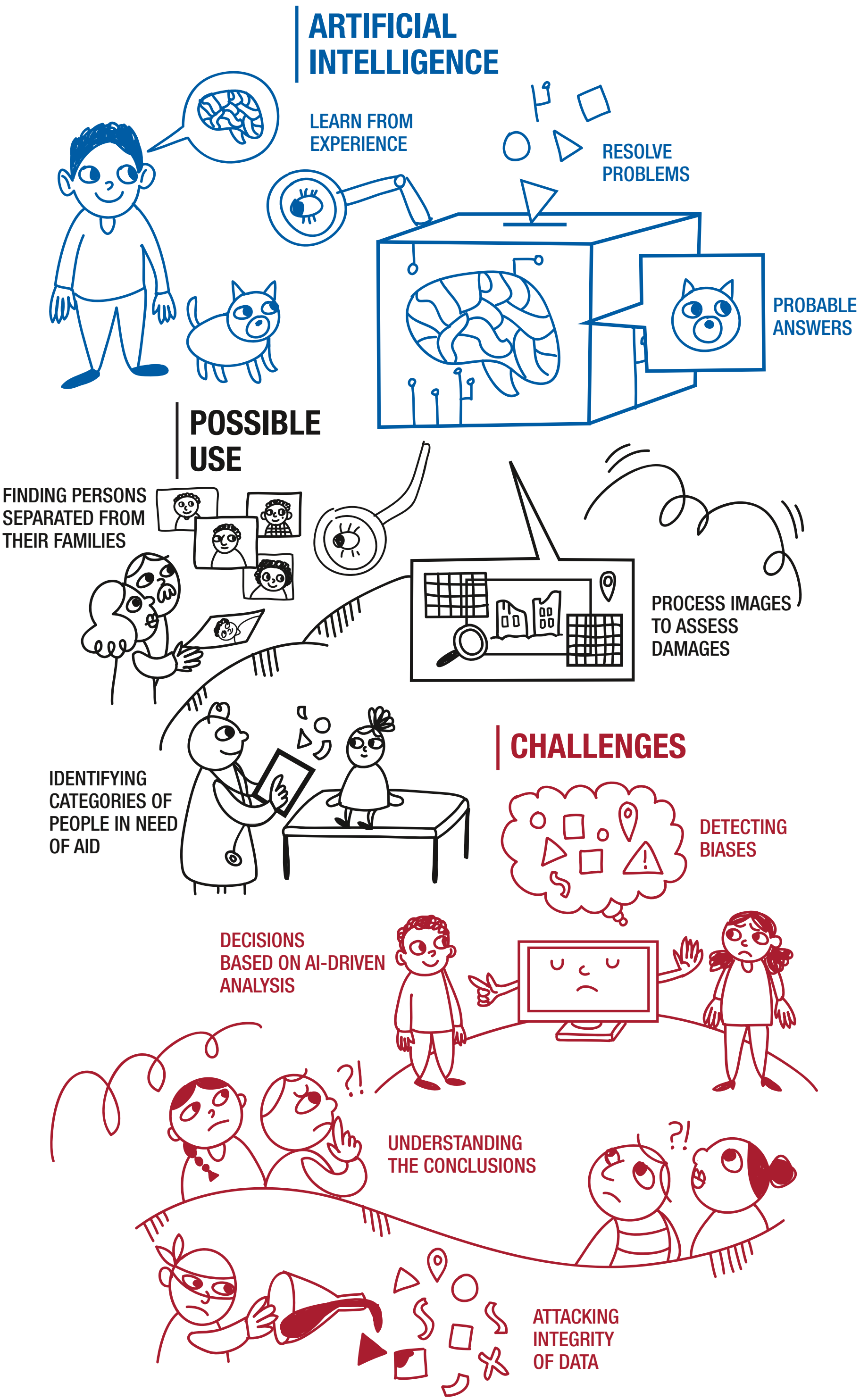


ARTIFICIAL INTELLIGENCE

LEARN FROM EXPERIENCE

RESOLVE PROBLEMS

PROBABLE ANSWERS

POSSIBLE USE

FINDING PERSONS SEPARATED FROM THEIR FAMILIES

PROCESS IMAGES TO ASSESS DAMAGES

IDENTIFYING CATEGORIES OF PEOPLE IN NEED OF AID

CHALLENGES

DETECTING BIASES

DECISIONS BASED ON AI-DRIVEN ANALYSIS

UNDERSTANDING THE CONCLUSIONS

ATTACKING INTEGRITY OF DATA

CHAPTER 17

ARTIFICIAL INTELLIGENCE

Alessandro Mantelero

17.1 INTRODUCTION[1]

This chapter explores the data protection challenges associated with the use of Artificial Intelligence systems in the humanitarian sector. The most relevant are some key elements of data Processing (such as the use of large data sets) and the purpose of such Processing, particularly as it concerns decision-making processes. The sections that follow first give a basic explanation of the technology in question, then identify the related data protection challenges and provide guidance for Humanitarian Organizations on how to address some of them.

17.1.1 WHAT ARTIFICIAL INTELLIGENCE IS AND HOW IT WORKS

While there is no single, universally accepted definition of the term, Artificial Intelligence is generally understood as "[a] set of sciences, theories, and techniques whose purpose is to reproduce by a machine the cognitive abilities of a human being".[2] In its current form, it aims to allow technology developers "to entrust a machine with complex tasks previously delegated to a human".[3]

Within the context of Artificial Intelligence, Machine Learning (ML) is one of the most relevant processes concerning the use of Personal Data in decision-making processes. This is a specific form of Artificial Intelligence defined as a set of algorithms that get better at completing a certain task over time, with input in the form of machine-readable data.[4] An ML algorithm receives more and more data representing the problem it is trying to solve and "learns" from such data. There are, however, other Artificial Intelligence techniques that are less reliant on data because they "learn" in different ways,[5] but, in recent years, Machine Learning has attracted the vast majority of Artificial Intelligence investment and is therefore the main reference for the considerations expressed in this chapter.

All forms of Artificial Intelligence share a common feature: they are not a set of instructions for a machine to complete a particular task, but rather a set of instructions for the machine to generate strategies or solutions to complete that task. There are different Artificial Intelligence techniques in existence, but for those relying on ML, it is possible to outline some common key elements as follows:

[1] This chapter builds on and revises two previous chapters of the second edition of this Handbook, on Big Data and AI respectively. The substance of these chapters was developed from a seminar developed with the contribution of the author during the Workshop on Artificial Intelligence/Machine Learning and Data Protection in Humanitarian Action, organised by VUB-Brussels Privacy Hub and the International Committee of the Red Cross in 2019.

[2] Council of Europe (CoE), "Glossary on Artificial Intelligence".

[3] Ibid.

[4] Tom M. Mitchell, *Machine Learning*, McGraw-Hill Series in Computer Science, McGraw-Hill, New York, 1997, 2.

[5] Examples of these methods include Bayesian networks and rule-based engines. These methods, however, are not addressed in this chapter.

1. Selected data sets relating to a certain field of investigation (e.g. human images for recognition or classification of persons) are presented to the system expecting that they contain specific patterns or similarities (training data).
2. Artificial Intelligence identifies these patterns by classifying/aggregating data according to relevant features present in the training data set.
3. This process generates a model that is able to recognize a pattern when new data are processed by it; these patterns support predictions or classifications related to the used data (e.g. mobile geolocated data to detect groups' mobility patterns).[6]

To understand the use cases of Artificial Intelligence, it is important to distinguish between three possible approaches to ML:

- **Supervised learning**: Training data are labelled by assigning a "class" to each piece of training data. For instance, images of animals are tagged with labels such as "dog", "cat" or "parrot" and fed into the system. Typically, the ultimate objective will be for the algorithm to be able to classify new (previously unseen) images into one of the learned classes. This type of learning can also be used, for example, to predict a value based on different parameters (or features), such as valuing a house based on the number of rooms, size and/or year of construction. In both cases, the objective is for the model to properly separate the data into their correct classes or evaluate correct values. In this process, data labelling is a crucial stage and requires field experts to identify key relevant elements, based on the data set and purpose of the analysis.
- **Unsupervised learning:** No labels are fed into the system, and Artificial Intelligence groups data based on similarities or patterns that it detects autonomously in the training data set. In this case, the classification is made by Artificial Intelligence during the learning process and no additional classes than those created by the ML process are possible.
- **Reinforcement learning:** This approach requires little training data. Instead, it relies on a method of reward and punishment, whereby "the system is given a 'reward' signal for when it accomplishes what the designer wants, or a step that advances the process toward the outcome the designer described. When the system does something wrong (fails to efficiently advance toward the desired outcome), it is simply not rewarded."[7]

Based on one of the methods described above,[8] it is possible to create static and dynamic models. Static models do not change over time and continue to apply the

[6] The Norwegian Data Protection Authority, *Artificial Intelligence and Privacy*, The Norwegian Data Protection Authority, Oslo, January 2018, 7: www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.

[7] Ibid., 18.

[8] This chapter does not address all possible Artificial Intelligence learning methods. For more information on methods not mentioned here (such as neural networks), see e.g.: Larry Hardesty, "Explained: Neural Networks", MIT News | Massachusetts Institute of Technology (blog), 14 April 2017: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414; Future of Privacy Forum, *The Privacy Expert's Guide to Artificial Intelligence and Machine Learning*, Future of Privacy Forum, Washington, DC, October 2018: https://fpf.org/wp-content/uploads/2018/10/FPF_Artificial-Intelligence_Digital.pdf.

model developed using the training data set. They give the developer better control over the model but prevent the adopted solution from improving over time. Dynamic models, on the other hand, are characterized by a kind of continuous learning, as they can use fresh data for improvements and changes (e.g. spam filter systems). This reduces control over the model development and may lead to unforeseen critical consequences in its outputs and expected behaviour.[9]

By nature, most of these Artificial Intelligence techniques rely on large-scale data sets, which are the main reason for their application and an inherent component of their functioning. Finding common patterns in a large amount of data – such as, for example, those produced at the national level on migration – might be hard for human experts. At the same time, the computer, statistical and mathematical tools used by Artificial Intelligence systems only work properly when applied to large data set minimizing outliers and other "noise" or disturbances.

Against this technology background, the progressive datafication of our society, due to the increasing availability of data produced by a variety of sources and the decreasing of the costs of sensors, IT devices/services and computing power, has made it possible to use Artificial Intelligence and to analyse large-scale data sets in all the fields of human activity, including Humanitarian Action.[10] A shift in the approach to social analysis followed the advent of so-called big data and Artificial Intelligence-based Data Analytics at the beginning of the new millennium. For the first time it was possible to combine very large volumes of diversely sourced information and analyse them, using mathematical algorithms at large scale or sophisticated computer-based tools (e.g. neural networks) to extract further information and make informed decisions.

However, this use of Artificial Intelligence for social analysis raises several questions and the risk of "algorithmic illusions".[11] Likewise, the way data collection is carried out, the design of the Artificial Intelligence model, the training data set used, and all potential errors or biases in this process, have an influence on the representation of human activities, relationships and profiles we use in Artificial Intelligence-supported Humanitarian Action tools.

[9] See e.g. the Microsoft Tay chatbot case: James Vincent, "Twitter Taught Microsoft's AI Chatbot to Be a Racist Asshole in Less than a Day", The Verge, 24 March 2016: www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

[10] See also United Nations Office for the Coordination of Humanitarian Affairs (OCHA), *Humanitarianism in the Age of Cyber-Warfare*. OCHA Policy and Studies series, 2014: www.unocha.org/publications/report/world/humanitarianism-age-cyber-warfare-towards-principled-and-secure-use-information.

[11] See also, on the use of Data Analytics in society, Alessandro Mantelero, "Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection", *Computer Law & Security Review*, Vol. 32, No. 2, 1 April 2016, pp. 238–255: www.sciencedirect.com/science/article/abs/pii/S0267364916300280?via%3Dihub.

Although the term "Artificial Intelligence" suggests that natural intelligence and artificial intelligence are similar, this is not the case. Artificial Intelligence is nothing more than a data-driven and mathematical form of information Processing; it is not able to think, elaborate concepts or develop theories of causality. Artificial Intelligence merely takes a path recognition approach to sort through very large amounts of data and infer new information and correlations. Data dependence and path dependence are therefore both the strength and the weakness of these systems, as well as the fact that AI-based solutions are designed to be applied serially and poor design therefore affects numerous people in the same or similar circumstances.

Finally, given the use of incredibly large data sets and complex Artificial Intelligence systems, the safeguarding role over decision making provided by human supervision may be very challenging and time-consuming, if not impossible in some cases.

In terms of its field-specific application, Artificial Intelligence and large data sets may be used for objectives such as identifying potential threats relevant to Humanitarian Action, enhancing preparedness, identifying individuals or categories of individuals in need, or predicting possible patterns of evolution of contagious diseases, conflicts, tensions and natural disasters. Data-driven technologies can significantly enhance the effectiveness of work carried out by Humanitarian Organizations, including mapping or identification of:

- patterns of events in Humanitarian Emergencies involving protected people in conflicts or other situations of violence;
- the spread of diseases or natural disasters, thus predicting possible developments and preparing to prevent damage;
- the epicentre of a crisis;
- safe routes;
- individual humanitarian incidents;
- vulnerable individuals or communities who are likely to require humanitarian response;
- matches in case of separated families in Humanitarian Emergencies.

Two broad categories of applications for the use of Artificial Intelligence-based solutions in Humanitarian Action can be identified:

(i) applications that recognize general patterns and predict trends;
(ii) applications aimed at identifying individuals or groups of individuals of relevance for Humanitarian Action.

In this context, the massive collection of data and the use of data-intensive applications based on personal information entails several risks. Not only might it lead to misleading and inaccurate results or decisions, but moreover the lack of accurate data protection-oriented design could lead to the development of invasive or disproportionate Artificial Intelligence systems, as well as the adoption of solutions affected by significant weaknesses that make it possible to reidentify individuals in poorly anonymized data sets, Data Breaches and other cybersecurity attacks.[12]

17.1.2 ARTIFICIAL INTELLIGENCE IN THE HUMANITARIAN SECTOR

Recent growth in available data and Processing power has greatly increased the number of Artificial Intelligence applications in everyday life: from virtual digital assistants to biometric recognition systems to unlock devices or allow access to buildings, from traffic management in smart cities to content moderation for online platforms, and in many other functionalities of online and offline products and services. Artificial Intelligence can also be applied to a wide variety of tasks traditionally performed by humans, such as medical diagnosis, image recognition and stock market prediction.

Regarding the application of Artificial Intelligence in the humanitarian sector, its ability to collect, process and analyse large data sets and to extract inferences and predictions to inform decision-making processes turns Artificial Intelligence into a valuable option to increase the efficiency and effectiveness of humanitarian work. This is evident, for example, in the use cases detailed below:

- **Reading public opinion.** In Uganda, the UN Global Pulse programme piloted “a toolkit that makes public radio broadcasts machine-readable through the use of speech recognition technology and translation tools that transform radio content into text”.[13] This tool, developed by the Pulse Lab Kampala, aims to identify trends among different population groups, particularly those in rural areas. The rationale behind the initiative is that these trends could then provide government and development partners with a better understanding of public opinion on the country’s development needs, which could then be taken into consideration when implementing development programmes.
- **Identifying and locating missing children.** It has been reported[14] that India’s National Tracking System for Missing & Vulnerable Children identified nearly 3,000 missing children within four days of launching a trial of a new facial recognition system that matches the faces of missing individuals with photographs of children living in children’s homes and orphanages.

[12] Marelli, “Defining the Cyber Perimeter”, April 2020, 367.

[13] Pulse Lab Kampala, “Making Ugandan Community Radio Machine-Readable Using Speech Recognition Technology”, UN Global Pulse (blog), 2016: www.unglobalpulse.org/project/making-ugandan-community-radio-machine-readable-using-speech-recognition-technology/.

[14] Anthony Cuthbertson, “Indian police trace 3,000 missing children in just four days using facial recognition technology”, *The Independent*, 24 April 2018: www.independent.co.uk/tech/india-police-missing-children-facial-recognition-tech-trace-find-reunite-a8320406.html; see also: PTI, “Delhi: facial recognition system helps trace 3,000 missing children in 4 days”, *The Times of India*, 22 April 2018: https://timesofindia.indiatimes.com/city/delhi/delhi-facial-recognition-system-helps-trace-3000-missing-children-in-4-days/articleshow/63870129.cms. For the system’s official website, see: https://trackthemissingchild.gov.in/trackchild/index.php/index.php.

- **Tracking attacks on civilians and human rights violations.** Amnesty International's Decode the Difference project[15] recruited volunteers to compare images of the same location at different time periods to identify damaged buildings, which could potentially demonstrate systematic attacks against civilians. In the future, the data could be used to train Machine Learning tools to analyse the images, thereby speeding up the process and increasing capacity.
- **Preventing and diagnosing disease.** "Since the 1990s, AI has been used to diagnose various types of diseases, such as cancer, multiple sclerosis, pancreatic disease and diabetes."[16] More recently, Microsoft's Project Premonition was developed to detect pathogens before they cause outbreaks. The project deploys robots that aim to monitor the presence of mosquitoes in an area, make predictions about their distribution and capture targeted species. Through Machine Learning techniques, the captured mosquitoes are searched for pathogens they may carry from animals they have bitten.[17]

When dealing with Artificial Intelligence-based projects, concerns may also be raised when applying basic data protection principles[18] in this context. Artificial Intelligence-based profiling and hidden nudging practices challenge the idea of freedom of choice based on the notion of Data Subjects' control over their information, and the widespread complexity and obscurity of Artificial Intelligence algorithms hamper the chances of obtaining real informed Consent and transparency requirements. Similar challenges relate to another key principle, data minimization, as big data and Machine Learning Artificial Intelligence algorithms rely on large amounts of data to produce useful results.[19]

[15] Amnesty International, "Amnesty Decoders | Join Decode Surveillance NYC", Amnesty International, accessed 17 March 2022: https://decoders.amnesty.org.

[16] Heather M. Roff, "Advancing Human Security through Artificial Intelligence", Chatham House – International Affairs Think Tank, 11 May 2017, 5: www.chathamhouse.org/2017/05/advancing-human-security-through-artificial-intelligence.

[17] Microsoft, "Microsoft Premonition", Microsoft Research (blog), accessed 21 March 2022: www.microsoft.com/en-us/research/product/microsoft-premonition/.

[18] See Chapter 2: Basic principles of data protection.

[19] See Council of Europe (CoE), *Guidelines on Artificial Intelligence and data protection | T-PD(2019)01*, Guideline (Strasbourg: Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108), 25 January 2019): rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8; Council of Europe (CoE), *Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data | T-PD(2017)01* (Strasbourg: Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108), 23 January 2017): rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806ebe7a. See also Alessandro Mantelero, *Beyond Data Human Rights, Ethical and Social Impact Assessment in AI*, 1st ed., Information Technology and Law Series, Springer, The Hague, 2022), chap. 1.

Before considering the specific issues related to Artificial Intelligence and large-scale data Processing, several specificities relating to data protection should be highlighted at the outset of this analysis:

- **Data sources.** First of all, it is important to identify the source of data. Much Artificial Intelligence-based data Processing undertaken by Humanitarian Organizations is based on publicly available data, such as information from government agencies or public records, social media networks, census data and other publicly available demographic and population surveys. In other cases, Humanitarian Organizations may partner with private enterprises such as telecommunications or infrastructure companies, Internet services, health-care providers or other commercial organizations to improve the humanitarian and disaster response.
- **Emergency response.** The outputs from Artificial Intelligence-ased data Processing can provide important benefits to Humanitarian Organizations. However, they may not always be used for an ongoing emergency or to address the vital interests of the people concerned: the exceptional, "outlier" circumstances where Humanitarian Organizations operate may become a limitation in predictive Machine Learning algorithms. Historical data sets and models in data-driven analyses, developed outside emergencies might find themselves scarcely able to cope due to outliers created in the extremely changeable circumstances of emergencies. Thus, it is important to consider Artificial Intelligence derived uniquely from Humanitarian Data since these models would integrate information learned during an emergency to support administrative work or to contribute to strategies to improve the response to future emergencies.
- **Accuracy.** Given the data-driven nature of Artificial Intelligence, the quality of the data used to train it significantly impacts both the development of the models and their performance. Here it is therefore crucial to verify that data used for training and running the Artificial Intelligence models are representative and accurate and do not contain any bias.[20]
- **Automated decisions.** Although in emergency situations automation can facilitate timely responses, it is important to be aware of the risks associated with a lack of human intervention and oversight, including in terms of ability to fully understand the complexity of the contextual background to prevent incorrect insights and decisions.
- **Reuse of data for other purposes.** The availability of large data sets often raises questions about the use of collected data for purposes other than those for which they were collected. This poses questions under Data Protection laws, which

[20] UN Global Pulse and Leiden University, "Big Data for Development and Humanitarian Action: Towards Responsible Governance", Global Pulse Privacy Advisory Group Meetings 2015–2016, December 2016: www.unglobalpulse.org/document/big-data-for-development-and-humanitarian-action-towards-responsible-governance. See also Mireille Hildebrandt, "The issue of bias: The framing powers of machine learning", in *Machine We Trust: Perspectives on Dependable AI*, ed. Marcello Pelillo and Teresa Scantamburlo, The MIT Press, Cambridge, MA, 2021, 44–59: https://dx.doi.org/10.2139/ssrn.3497597.

generally require that personal data be collected for specific purposes and processed for such purposes or for compatible purposes only, and not reused for other purposes without the Consent of the person concerned or another legal basis (see Section 17.2.1 – Legal bases for Personal Data Processing).

- **The sensitivity of data output created by Personal Data Processing in humanitarian situations.** It is important to understand that publicly available data, such as data on social media networks, mobility data or data generated by mobile phone connections, may generally be considered non-Sensitive Data but may generate Sensitive Data in different contests and mainly in a humanitarian situation. This can occur when the Processing of non-Sensitive Data enables the profiling of individuals that could be subjected to discrimination or repression, such as, for example, potential victims, people affiliated with a particular group in a situation of violence, or persons suffering from a particular illness. In these cases, specific computing techniques, such as *differential privacy*,[21] can be a valuable way to protect individual and group privacy while allowing access to data.[22]
- **Anonymization.** There may be doubts about the effectiveness of Anonymization of Personal Data and the possibility of Reidentification in Artificial Intelligence-based operations, regardless of whether for humanitarian or other purposes. Again, privacy-enhancing technologies, such as *synthetic data*,[23] can complement Anonymization attempts to provide higher protection and prevent Reidentification.[24]
- **Regulatory fragmentation.** While many states have enacted data protection laws and many Humanitarian Organizations have already implemented data protection policies and guidelines, the question of how specifically data and Artificial Intelligence-based data Processing are regulated across borders in times of humanitarian crises remains open.[25]

It is important to stress that when Artificial Intelligence is used for Humanitarian Action, the implications for individuals may be much more serious than in other

[21] Cynthia Dwork, “Differential Privacy”, in Henk C. A. van Tilborg and Sushil Jajodia (eds), *Encyclopedia of Cryptography and Security*, Springer US, Boston, MA, 2011, 338–340: https://doi.org/10.1007/978-1-4419-5906-5_752.

[22] Data smoothing means removing noise from a data set so that important patterns stand out.

[23] Synthetic data is information generated by algorithms that is not real-world data but reflects real-world data, mathematically or statistically. See European Data Protection Supervisor (EDPS), “IPEN Webinar 2021 – ‘Synthetic Data: What Use Cases as a Privacy Enhancing Technology?’”, EDPS, 16 June 2021: www.edps.europa.eu/data-protection/our-work/ipen/ipen-webinar-2021-synthetic-data-what-use-cases-privacy-enhancing_en.

[24] Prokopios Drogkaris and Monika Adamczyk (eds.), *Data Protection Engineering – From Theory to Practice*, European Union Agency for Cybersecurity (ENISA), 27 January 2022: www.enisa.europa.eu/publications/data-protection-engineering.

[25] UN Global Pulse and Leiden University, *Big Data for Development and Humanitarian Action*, 7–9. See also Júlia Zomignani Barboza, Lina Jasmontaitė-Zaniewicz and Laurence Diver, “Aid and AI: The challenge of reconciling humanitarian principles and data protection”, in *Privacy and Identity Management. Data for Better Living: AI and Privacy*, IFIP International Summer School on Privacy and Identity Management, Springer, Cham, 2020, 161–176: https://doi.org/10.1007/978-3-030-42504-3_11.

contexts. Humanitarian Organizations should therefore consider whether any data they release or information they provide using data-intensive Artificial Intelligence systems can be used, even in an aggregated form, to target the people they seek to protect. Furthermore, information on "invisible populations" can be extracted indirectly using data on different groups related to them, with potential implications in terms of discrimination or actions against minorities, even more so in case of conflicts. It is important, therefore, always to keep in mind the "big picture" of the potential implications of using data-intensive Artificial Intelligence systems in a context characterized by reduced protection systems and heightened vulnerabilities.

EXAMPLE:
Authorities might use public or published findings based on the extraction and analysis of tweets and other material on social media networks to locate the epicentre and flows of public demonstrations, and to avoid loss of human life. However, these same findings might then be used by the same authorities to identify individuals who took part in such public demonstrations (or who did not), which can have severe consequences for the identified groups of individuals.

Artificial Intelligence may involve Processing scenarios such as the following:

EXAMPLE 1: the extraction and analysis of public communications through social media, search engines or telecommunications services, as well as news sources. This can help demonstrate how methods including sentiment analysis, topic classification and network analysis can be used to support public health workers and communication campaigns.

EXAMPLE 2: the development of interactive data visualization tools during a humanitarian incident. This can help demonstrate how communications signals or satellite data could support emergency response management.

EXAMPLE 3: Analysis of messages received through a Humanitarian Organization's citizen reporting platform.

EXAMPLE 4: Analysis of social media, mobile phone network metadata and credit card data to identify individuals likely to be at risk of enforced disappearance or to locate persons unaccounted for.

Focusing on the large-scale data sets potentially used by Artificial Intelligence, the following may be relevant:

- **accessible data sets**: i.e. data sets that are already publicly available, such as public records released by governments or information people have intentionally made public in the media or on the Internet, including through social media;
- **data sets held by Humanitarian Organizations**: e.g. lists of distribution beneficiaries, patients, protected individuals, individuals reporting violations of international humanitarian law/human rights;

- **data sets held by private Third Parties**: e.g. mobile telecommunications, Internet service, banking and financial providers, financial transactions data, remote sensor data, whether aggregated/pseudonymized or not;
- **a combination or aggregation of data sets** of Humanitarian Organizations, authorities and/or corporate entities (including the organizations mentioned above).

Humanitarian Organizations may play the following roles in data Processing:

- process data held for the purposes of their respective organizations, in their capacity as Data Controllers or Joint Controllers (when determining the purposes and means of Processing jointly with other Humanitarian Organizations, public authorities and/or commercial entities);
- employ Third Parties who process data on behalf of the organization (e.g. commercial entities that use Artificial Intelligence for predictive analyses on the data held by the Humanitarian Organization and for the purposes of this organization) and act as Data Processors;
- require commercial entities that are and remain the Data Controller to carry out analyses on data for humanitarian purposes and to provide conclusions/findings to the Humanitarian Organization. Such conclusions may relate to aggregated/pseudonymized data, or data identifying individuals of possible relevance to Humanitarian Action.

17.1.3 CHALLENGES AND RISKS OF USING ARTIFICIAL INTELLIGENCE

Despite their potential, Artificial Intelligence applications carry challenges and risks. Besides data protection concerns,[26] all the above-mentioned use cases also present practical implementation challenges. For example, Artificial Intelligence-based image recognition software used to identify missing people may provide too many false positives. These false matches could not only create confusion among case workers, but also potentially give false hope to families. Other systems could be more accurate but potentially miss positive matches (known as false negatives). While false negatives may not be much of an issue in commercial applications, they can have important consequences in the humanitarian sector. If an organization misidentifies a child who has lost contact with their parents, this can cause harm to the entire family.

Artificial Intelligence can also pose risks to affected people. For instance, if Artificial Intelligence is used to identify the right target population for a particular humanitarian programme, and the solution does not make a correct identification, people who

[26] See also Anne Meuwese, “Regulating algorithmic decision-making one case at the time: A note on the Dutch ‘SyRI’ judgment”, *European Review of Digital Administration & Law*, Vol. 1, No. 1, 2020, pp. 209–211.

would otherwise be entitled to participate in the programme could be excluded. This has happened in practice in Sweden, where thousands of unemployed people were wrongly denied benefits by a government system that used Artificial Intelligence.[27]

Since most Humanitarian Organizations will acquire off-the-shelf solutions rather than developing their own models, there is a not-insignificant risk that algorithms could deliver unexpected or unreasonable results. This also highlights the risk of decontextualization when choosing off-the-shelf Artificial Intelligence models – where models originally used for one purpose are then reused in a different context and for a different purpose[28] – or when using models trained on historical data from a different population.[29]

In addition, vendor lock-in poses a risk because switching solutions may be costly. Organizations could also be targeted by commercial ventures that are primarily interested in gaining access to and exploiting the large data sets they hold, sometimes at great risk to the individuals and communities to whom the data belong.

Bias poses another risk to the effectiveness of Artificial Intelligence, especially in specific humanitarian contexts where it is important to use data sets fit for the intended goal. As with many other technologies, the concept of “garbage in, garbage out”[30] also applies to Artificial Intelligence, and using unfit, inaccurate or irrelevant data may affect the accuracy of the solution. This is particularly challenging for a Humanitarian Organization, as off-the-shelf algorithms will extremely rarely fit their contexts. For instance, if a Humanitarian Organization wants to develop facial recognition software to help find missing people, the training data sets will need to be sufficiently broad to ensure that racial variations in physical features are integrated to maximize the precision of the matching function.

[27] Tom Willis, “Sweden: Rogue Algorithm Stops Welfare Payments for up to 70,000 Unemployed”, AlgorithmWatch, 25 February 2019: https://algorithmwatch.org/en/rogue-algorithm-in-sweden-stops-welfare-payments.

[28] See Robyn Caplan et al., “Algorithmic Accountability: A Primer”, Data & Society, 18 April 2018, 7: https://datasociety.net/library/algorithmic-accountability-a-primer, citing the case of the PredPol algorithm, originally designed to predict earthquakes and later used to identify crime hotspots and assign police.

[29] See Council of Europe (CoE), *Guidelines on Artificial Intelligence and data protection | T-PD(2019)01*. See also Council of Europe (CoE), *Artificial Intelligence and Data Protection: Challenges and Possible Remedies | T-PD(2018)09Rev*, report on Artificial Intelligence (Strasbourg: Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108), 25 January 2019): https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6.

[30] According to the free online dictionary of computing (http://foldoc.org), the concept of garbage in, garbage out relates to the fact that “computers, unlike humans, will unquestioningly process nonsensical input data and produce nonsensical output”. The term is also used to refer to “failures in human decision-making due to faulty, incomplete, or imprecise data”.

Processing Personal Data using Artificial Intelligence also presents major challenges for Personal Data protection. When Processing large data sets for purposes other than those for which they were collected, there is a risk of violating basic notions of data protection, including purpose limitation, data minimization or data retention (i.e. keeping data only as long as necessary to fulfil the purposes of data collection).[31] In essence, large-scale data analysis thrives in open and unrestricted Processing environments while, on the other hand, Personal Data protection favours limited and well-defined Processing. Data protection thus needs to be applied in an innovative way to these technologies.[32]

The fundamental principles of data protection must be respected while performing Artificial Intelligence-based data Processing. These principles include (i) fairness and lawfulness of the Processing; (ii) transparency; (iii) purpose limitation; (iv) data minimization; (v) data quality. While some of these principles are compatible with the nature of Artificial Intelligence applications, others raise questions or conflicts.[33] Consequently, Humanitarian Organizations must be particularly careful when applying them in practice.[34]

17.2 APPLICATION OF BASIC DATA PROTECTION PRINCIPLES

Solutions that integrate or use Artificial Intelligence process large amounts of data – both personal and non-personal – in order to function properly. In this regard, it is crucial to consider that these applications can infer Personal Data from non-personal information or anonymized data. This is because Artificial Intelligence solutions are increasingly capable "of linking data or recognizing patterns of data [that] may render non-personal data identifiable".[35] This means that Artificial Intelligence can also reidentify data provided, for example, by a variety of sensors and smart devices.

[31] See Chapter 2 – Basic principles of data protection. See also: Council of Europe (CoE), *Artificial Intelligence and Data Protection: Challenges and Possible Remedies | T-PD(2018)09Rev*.

[32] See Council of Europe (CoE), *Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data | T-PD(2017)01*; Alessandro Mantelero, "Regulating Big Data: The guidelines of the Council of Europe in the context of the European Data Protection Framework", *Computer Law & Security Review*, Vol. 33, No. 5, October 2017, pp. 584–602: www.sciencedirect.com/science/article/abs/pii/S0267364917301644?via%3Dihub; UN Global Pulse, *Guidance note on Big Data for achievement of the 2030 Agenda*, 19 August 2019: www.unglobalpulse.org/policy/privacy-and-data-protection-principles; European Data Protection Supervisor (EDPS), *Opinion 7/2015: Meeting the Challenges of Big Data*, Opinion, EDPS, Brussels, 19 November 2015, 4: www.edps.europa.eu/data-protection/our-work/publications/opinions/meeting-challenges-big-data_en.

[33] See also Mantelero, *Beyond Data Human Rights, Ethical and Social Impact Assessment in AI*, chap. 1.

[34] The discussion on data protection in this chapter builds on the principles set out in Part I and examines them in greater detail.

[35] Centre for Information Policy Leadership, *Artificial Intelligence and Data Protection in Tension*, Artificial Intelligence and Data Protection: Delivering Sustainable AI Accountability in Practice,

An assessment of the risks of Reidentification should therefore be carried out and, when possible, the Data Subject or relevant stakeholders be informed of the results of this assessment. If there is a strong possibility of Reidentification, the analysis should not be performed, or the methodology should be adjusted.

For these reasons, the use of Anonymization as an "exit strategy" with respect to data protection obligations is not always effective. Moreover, anonymous, or anonymized data may also present technical challenges as the capacity to process may be hindered during Processing.

In addition, the accuracy of Artificial Intelligence outputs when Processing anonymized or aggregated data should be assessed. The methods and level of Anonymization or aggregation should therefore be carefully selected to minimize not only the risks of Reidentification but also to ensure that the data maintain an adequate level of quality to achieve credible results.

17.2.1 LEGAL BASES FOR PERSONAL DATA PROCESSING

When carrying out Artificial Intelligence-driven Processing operations, Humanitarian Organizations may rely on one or more of the following legal bases:[36]

- the vital interest of the Data Subject or of another person;
- the public interest, in particular based on an Organization's mandate under national or international law;
- the informed Consent of the Data Subject;
- a legitimate interest of the organization;
- the performance of a contract;
- compliance with a legal obligation.

However, the specific nature of Artificial Intelligence applications and related data Processing poses some challenges to this traditional framework, mainly in the case of individual Consent to data Processing and secondary use of collected data (i.e. data originally collected for a specific purpose and then reused for a different one, as is often the case in Artificial Intelligence given the large-scale data sets needed).

As pointed out in literature, the effectiveness of Data Subjects' Consent as a legal basis has been weakened by lengthy and technical data Processing notices, social and technical lock-ins, obscure interface design, and lack of awareness on the part of the Data Subject.[37] These developments are even more relevant in the context of

10 October 2018, 11: www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_ai_first_report_-_artificial_intelligence_and_data_protection_in_te....pdf.

[36] See Chapter 3: Legal bases for Personal Data Processing.

[37] For a broader analysis and refences see Alessandro Mantelero, "The future of consumer data protection in the E.U. Re-thinking the 'notice and consent' paradigm in the new era of predictive analytics",

Humanitarian Action, when Data Subjects already experience imbalances of power and other contextual needs that hamper their effective self-determination.

Moreover, Artificial Intelligence-based profiling and hidden nudging practices challenge both the idea of freedom of choice based on contractual agreement and the notion of Data Subjects' control over their personal information. Finally, the frequent complexity and obscurity of Artificial Intelligence algorithms hamper the possibilities of obtaining truly informed Consent.

Legal scholars have addressed these issues by emphasizing the role of transparency,[38] risk assessment[39] and more flexible forms of Consent, such as broad Consent[40] or dynamic Consent.[41] Although none of these solutions solve the problems affecting individual Consent, in certain contexts they may, whether alone or combined, reinforce self-determination.

Notwithstanding these unresolved critical issues in terms of theoretical framework and regulatory instruments, Consent can be a legitimate ground for the Processing data collected by a Humanitarian Organization, but also for the reuse of data collected by Third Parties for different purposes. An example in this sense is the Data Analytics offered by social media networks or mobile phone operators to assist Humanitarian Organizations which could, in some cases, be based on Consent. In such cases, the social media platform or mobile operator in question can inform Data Subjects of the intended Processing by means of a pop-up window or text message with the relevant information and provide a Consent request.

Computer Law & Security Review, Vol. 30, No. 6, 1 December 2014, pp. 643–660: https://doi.org/10.1016/j.clsr.2014.09.004.

[38] See e.g.: Lilian Edwards and Michael Veale, "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for", *Duke Law & Technology Review*, 16, 2018 2017, pp. 18–84; Andrew Selbst and Julia Powles, "'Meaningful information' and the right to explanation", *International Data Privacy Law*, Vol. 7, No. 4, 19 December 2017, pp. 233–242: doi.org/10.1093/idpl/ipx022: https://proceedings.mlr.press/v81/selbst18a.html; Sandra Wachter, Brent Mittelstadt and Luciano Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation", *International Data Privacy Law*, Vol. 7, No. 2, 1 May 2017, pp. 76–99: https://doi.org/10.1093/idpl/ipx005.

[39] See Council of Europe (CoE), *Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data | T-PD(2017)01*; Mantelero, "Regulating Big Data:The guidelines of the Council of Europe in the Context of the European Data Protection Framework"; UN Global Pulse, *Guidance note on Big Data for achievement of the 2030 Agenda*; European Data Protection Supervisor (EDPS), *Opinion 7/2015: Meeting the Challenges of Big Data*.

[40] Mark Sheehan, "Can broad consent be informed consent?", *Public Health Ethics*, Vol. 4, No. 3, 1 November 2011, pp. 226–235: https://doi.org/10.1093/phe/phr020.

[41] Jane Kaye et al., "Dynamic consent: A patient interface for twenty-first century research networks", *European Journal of Human Genetics*, Vol. 23, No. 2, February 2015, pp. 141–146: doi.org/10.1038/ejhg.2014.71.

In order to ensure that the Data Subject receives adequate information before giving Consent, such information should include the outcome of the DPIA (if carried out)[42] and could also be provided via an interface that simulates the effects of the use of data and their potential impact on the Data Subject, in a learn-from-experience approach.[43] Data Controllers should provide Data Subjects with easy and user-friendly technical ways to withdraw their Consent and react to data Processing incompatible with the initial purposes.[44]

It is important to assess the validity of Consent even when adequate information has been provided to the Data Subjects at the time of collection and the purpose of Further Processing is compatible. This assessment should take into account the level of literacy of the Data Subject as well as the risks and harms to the Data Subjects for the Processing of their data.[45]

Without the Consent of the Data Subject, Personal Data can be processed in the vital interest of the Data Subject or of another person, i.e. where data Processing is necessary in order to protect an interest essential in the life, integrity, health, dignity and safety of the Data Subject or that of another person or group of people. Furthermore, additional legal bases, such as public interest, the legitimate interest of the organization and performance of a contract or compliance with a legal obligation may also be grounds for data Processing.

Regarding the use of vital interest as a legal basis for emergency work of Humanitarian Organizations in armed conflicts and other situations of violence, there are several cases where the Processing of data by Humanitarian Organizations is presumed to be in the vital interest of the Data Subject or another person (e.g. if data are processed in cases of Sought Persons, or if there are imminent threats against the physical and mental integrity of the persons concerned). However, the condition of vital interest may not be met when data Processing is carried out in a non-emergency situation, for instance for administrative purposes.

Humanitarian Organizations should carefully consider the existence of important public interests, which are sufficiently closely linked to Artificial Intelligence-based operations envisaged, to be used as a legal basis for Processing Personal Data. The public interest could be the appropriate legal basis for data Processing where a mandate to carry out a Humanitarian Action is established in national, regional or

[42] See Section 17.6: Data Protection Impact Assessment and Human Rights Impact Assessment.

[43] Council of Europe (CoE), *Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data | T-PD(2017)01*.

[44] Ibid.

[45] UN Global Pulse, “Tools: Risks, Harms and Benefits Assessment”, updated 2020: www.unglobalpulse.org/policy/risk-assessment.

international law and where no Consent was obtained and no emergency exists that could invoke vital interest as a legal basis.

Humanitarian Organizations should be aware that public interest as a legal basis for Personal Data Processing is not transferable, because it is specific to the Organization's mandate under national or international law. The conditions (if any) under which a Third Party may undertake the data analysis, including using Artificial Intelligence, on behalf of the Organization or that are applicable to International Data Sharing need to be examined separately.

Humanitarian Organizations may also process Personal Data where this is in their legitimate interest, provided that this interest is not overridden by the fundamental rights and freedoms of the Data Subject. Such legitimate interests may include Processing necessary to make their operations more effective and efficient, including facilitating logistics to enable pre-deployment of aid and staff in anticipation of Humanitarian Emergencies, where such insights could be obtained from data analysis. The use of Artificial Intelligence for administrative purposes may also fall under this category.

17.2.2 PURPOSE LIMITATION AND FURTHER PROCESSING

One of the most significant challenges in using Artificial Intelligence for humanitarian purposes is that Artificial Intelligence operations are very likely to be run on existing data sets, previously collected by the Humanitarian Organization or by Third Parties for a different purpose. The key question is, therefore, to determine whether the envisaged analysis is compatible with the original purpose of collection. If so, Artificial Intelligence operations can be carried out under the existing legal basis. If not, a new legal basis for Further Processing must be found.

In addition, applying the purpose limitation principle[46] to Artificial Intelligence may be challenging because these technologies have the capacity to process data in ways that were not originally planned, and are used to identify new patterns and inferences which are, by their nature, unknown and unexpected.

EXAMPLE:

In 2012, researchers found that when Artificial Intelligence algorithms analysed a person's Facebook "likes", with no further information from that person, the solutions could "automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views,

[46] See Section 2.5.2 – The purpose limitation principle.

personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender".[47] More specifically, the solution correctly discriminated "between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrat and Republican in 85% of cases".[48] In this particular case, the solution was being asked to make these correlations. Yet in other situations, Artificial Intelligence solutions may draw such inferences on their own and reveal sensitive information about a person even when that was not the developer's intention.

As discussed in Chapter 2: Basic principles of data protection, at the time of collecting data the Humanitarian Organization concerned must determine and set out the specific purpose(s) for which data are processed. The specific purpose(s) should be explicit and legitimate and could include anything from restoring family links, to protecting individuals in detention, forensic activities or protecting water and habitat. The purpose of any planned analytics should be specified at the outset of data collection, and when new purposes are added this must be consistent with the data protection requirements in terms of compatible purposes and legal grounds.

Artificial Intelligence – in a similar way to big data[49] – represents a challenge for the application of the purpose limitation principle. On the one hand, analytics make it hard to identify the specific purpose of data Processing at the time of data collection and, on the other hand, Machine Learning algorithms (whose purposes are necessarily specified) may not anticipate and explain how these purposes are to be achieved. In both cases therefore transparency on the purpose and methods of data Processing may remain limited.

In addition, the purpose limitation principle should also be considered with regard to the data sets used and potential unwanted outcomes. If it is foreseen that the solution may process Personal Data in ways that are incompatible with the defined purpose or that it will reveal information or make predictions that are not desired, these factors should be taken into account when choosing the training data set and developing the model.

In these large-scale data-intensive applications, it is common to carry out Processing operations that require the data to be processed for purposes other than those for which they were initially collected. In this case of secondary use of data,

[47] Michal Kosinski, David Stillwell and Thore Graepel, "Private traits and attributes are predictable from digital records of human behavior", *Proceedings of the National Academy of Sciences*, Vol. 110, No. 15, 11 March 2013, p. 1: https://doi.org/10.1073/pnas.1218772110.

[48] Ibid.

[49] See also The Norwegian Data Protection Authority, *Artificial Intelligence and Privacy*.

Humanitarian Organizations may therefore assess whether Further Processing is compatible with the purposes initially specified at the time of data collection, including where the Processing is necessary for historical, statistical or scientific purposes.[50]

In order to establish whether these operations can be considered Further Processing that is compatible with the purpose for which the data were initially collected, attention should be given to the following factors:

- any link between the purposes for which the data were collected and the purposes of the intended Further Processing;
- the situation in which the Personal Data were collected and, in particular, the relationship between Data Subjects and the Data Controller, and possible expectations of the Data Subjects;
- the nature of the Personal Data;
- the possible consequences of the intended Further Processing for Data Subjects;
- the existence of appropriate safeguards.

Based on these factors, it is possible that in several cases different humanitarian purposes are linked and considered compatible with each other. Compatibility depends on the circumstances of the case and Further Processing would not be compatible if new risks arise, or if the risks for the Data Subject outweigh the benefits of Further Processing. Further Processing would also not be compatible where Processing is potentially detrimental to the interests of the Data Subject or his/her family, in particular when there is a risk that the Processing might threaten their life, integrity, dignity, psychological or physical safety, freedom or reputation. This includes consequences such as harassment or persecution by authorities or Third Parties, judicial prosecution, social and private problems, restriction of freedom, and psychological suffering.

It should also be highlighted that some data protection regulations, such as the EU GDPR, pose restrictions to secondary uses of Personal Data but adopt specific derogations for public interest purposes, which include humanitarian purposes. In cases, where Third Party data are processed for purposes that go beyond those for which they were originally collected due to the humanitarian value in the use of the data sets, humanitarian purposes should not expose the Data Subjects to new risks or harm.

EXAMPLE 1: Data sets collected by a Humanitarian Organization while dealing with an incident, for instance in order to distribute aid, may be used at a later stage for the

[50] See Subsection 2.5.2.1 – Further Processing.

purpose of understanding patterns of displacement and pre-deploying aid in subsequent Humanitarian Emergencies.

EXAMPLE 2: Data sets collected by a telecommunications provider in the course of providing its services to its subscribers may not be used without these subscribers' Consent in Data Analytics Processing by Humanitarian Organizations, if it can result in such individuals being profiled as potential bearers of a disease, with consequent restrictions on movement imposed by authorities. In these cases, Humanitarian Organizations and their Third Party counterparts should consider whether mitigating measures, such as data aggregation, would be sufficient to remove the risk identified.

17.2.3 FAIR AND LAWFUL PROCESSING

As is always the case with Personal Data Processing, if Personal Data will be processed within the Artificial Intelligence solution or as part of its training, a lawful process requires a legitimate legal basis for the Processing to take place. Chapter 3: Legal bases for Personal Data Processing, outlines different legal grounds and points out the limitations of using Consent as a legal basis in Humanitarian Action. Limitations to the use of Consent, in particular the possibility of withdrawing it, are also relevant to the development and improvement of Artificial Intelligence solutions.[51]

When a Humanitarian Organization develops an Artificial Intelligence-based solution, it should identify an appropriate legal basis to process Personal Data to train the algorithm to achieve a clearly defined purpose. A legal basis should also be defined for the Processing of new Personal Data to fulfil the intended objective once the system has been trained. Lastly, the organization should also identify a legal basis for Processing data to improve the model, in the case of dynamic models.

With dynamic models, including off-the-shelf solutions developed by technology companies, it is important to remember that all data fed into the system during development and application will be used to improve it. This may pose further challenges to the use of Consent, since beneficiaries might agree to having their Personal Data processed for a particular humanitarian purpose, but may not expect it to be used for the development of the Artificial Intelligence solution.[52] In such cases, if the identified legal basis for Processing is Consent, the Data Subjects should be informed, in an easy-to-understand manner, of the reasons why their data are requested, what they will be used for, and how they will influence the solution. They should also be informed of potential risks, such as Reidentification by the solution or the fact that their data could be accessed during a malicious attack.

[51] See also above Section 17.2.1 – Legal bases for Personal Data Processing.

[52] Future of Privacy Forum, *The Privacy Expert's Guide to Artificial Intelligence and Machine Learning*, 8.

In light of the above, Consent may not always be an appropriate legal basis for the use of Artificial Intelligence in the humanitarian sector. While the delivery of aid or life-saving services may mean that vital interest[53] or public interest[54] can be considered legitimate legal bases to justify the Processing of Personal Data, the development of Artificial Intelligence solutions sometimes may not. To determine whether the improvement of Artificial Intelligence solutions is acceptable under the chosen legal basis, an organization should consider whether the Further Processing for the improvement of the solution is compatible with the initial purpose for which it collected the Personal Data.

The principle of fairness[55] requires that all Processing activities respect Data Subjects' interests, and that Data Controllers take action to prevent arbitrary discrimination against individuals.[56] The issue of discriminatory bias in Artificial Intelligence is widely recognized and debated.[57]

EXAMPLE:
In a well-known example, an Artificial Intelligence solution was developed in the United States to predict reoffending rates in criminal cases, in order to help judges decide whether or not to grant bail to convicted offenders. The solution incorrectly rated black defendants as being almost twice as likely to reoffend as white defendants.[58]

To minimize the risk of discriminatory bias, it is recommended that Artificial Intelligence developers "adopt a human rights by-design approach and avoid any potential biases, including unintentional or hidden, and the risk of discrimination or other adverse impacts on the human rights and fundamental freedoms of data subjects".[59]

Bias in Artificial Intelligence solutions may stem from the use of biased data sets as training data, from systemic biases in society, or even from developers deciding

[53] See Section 3.3 – Vital interest.

[54] See Section 3.4 – Important grounds of public interest.

[55] See Section 2.5.1 – The principle of the fairness and lawfulness of Processing.

[56] The Norwegian Data Protection Authority, *Artificial Intelligence and Privacy*, 16.

[57] Sandra Wachter, Brent Mittelstadt and Chris Russell, "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI", *Computer Law & Security Review*, Vol. 41 (2021), 105567.

[58] Julia Angwin et al., "Machine Bias", ProPublica, 23 May 2016: www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token = p-v0T1xjfOJ8jrJHzc08UxDKSQrKgWJk.

[59] Council of Europe (CoE), *Artificial Intelligence and Data Protection: Challenges and Possible Remedies | T-PD(2018)09Rev*, 2.

which features to assign more value to in each data set. Moreover, when there are historical biases in society, it may be difficult to find unbiased data to train Artificial Intelligence or it is necessary to "clean" or normalize the data sets or adopt alternative solutions such as debiased synthetic data.

More generally, to prevent bias, a model must be trained with relevant and correct data and must also learn which features to emphasize. Depending on the case, when there is a risk of arbitrary discrimination, information related to racial or ethnic origin, political opinion, religious and philosophical beliefs, sexual orientation or any other information that could be grounds for discrimination may not be processed or may be protected in a way that does not emphasize them leading to discrimination.[60]

The training data must also be fit for the purpose of the Artificial Intelligence solution. In other words, the selected data must be relevant to the task, and constant checks and updates will be required to identify inaccurate and/or corrupt data and remove them from the training data set. New data may also be added to avoid bias. It is therefore important that Humanitarian Organizations work with developers to ensure that the solution they acquire or develop is applicable or suited to the organization's needs in a particular context.

The fact that Artificial Intelligence models should not emphasize such categories of data does not mean, however, that suppressing them from the data set will necessarily eliminate the risk of bias. The system could correlate other features such as race or gender, and the model may learn to be biased based on those correlated features, which are known in this context as "proxies".[61] Moreover, since the main discriminatory feature has been removed from the data set, it might be more difficult to detect and correct the bias.

EXAMPLE:
A separate study looking at the US predictive solution discussed earlier found in almost 70 per cent of cases that the algorithm made a correct reoffending prediction despite its clear bias. In this second study, however, race was not included in the data set, highlighting "the challenge of finding a model that doesn't create a proxy for race (or other eliminated factor) – such as poverty, joblessness, and social marginalization".[62]

[60] The Norwegian Data Protection Authority, *Artificial Intelligence and Privacy*, 16.

[61] Centre for Information Policy Leadership, *Artificial Intelligence and Data Protection in Tension*, 14.

[62] Future of Privacy Forum, *The Privacy Expert's Guide to Artificial Intelligence and Machine Learning*, 15.

For this reason, when choosing the training data set, an Artificial Intelligence developer – whether acting as an independent Data Controller, a Data Processor, or a joint Controller with a Humanitarian Organization – needs to assess the quality, nature and origin of the Personal Data used, and consider the potential risks to individuals and groups of using decontextualized data to create decontextualized models.[63] One way to achieve this is for Data Controllers to include, in the continuous DPIA process (see Section 17.2 – Application of basic data protection principles), "frequent assessments on the datasets they process to check for any bias", and to "develop ways to address any prejudicial elements, including any over-reliance on correlations".[64] Not taking such measures has both legal and ethical implications.

In addition, Artificial Intelligence deals with possible correlations and therefore raises concerns about data selection, representation and population estimates. Researchers should take care to understand the representativeness of the data used and report potential biases. Moreover, policymakers should be aware of potential biases and account for them when making decisions, as inaccurate and biased data could lead to harmful and unfair policy decisions.

Finally, we could also identify a procedural component of fairness, requiring that any employees, contractors or other parties involved in data Processing undergo training to educate them about these risks and the steps to be taken to mitigate them.

17.2.4 TRANSPARENCY

Alongside fairness, transparency is another crucial aspect of data protection. According to this principle, the Processing of Personal Data must be transparent[65] for the Data Subjects involved, who should receive key information concerning the Processing when their data are collected.[66]

Transparency also contributes to the application of the fairness requirement in data protection. Given the complexity of the Processing, transparency on its methodology (including where possible the algorithm) is very important, so that the rigour of the approach can be independently assessed (beyond the Data Subjects' right of information[67]) and is the main requirement to perform a meaningful risk analysis.

Transparency, however, can be a challenging principle to apply when it comes to Artificial Intelligence, since these solutions are based on advanced technology that

[63] Council of Europe (CoE), *Artificial Intelligence and Data Protection: Challenges and Possible Remedies | T-PD(2018)09Rev*, 2.

[64] Article 29 Data Protection Working Party, *Guidelines on automated individual decision-making*.

[65] See Section 2.5.1 – The principle of the fairness and lawfulness of Processing.

[66] See Section 2.10 – Information.

[67] See Section 17.3 – Rights of Data Subjects, and Section 17.5 – International Data Sharing.

can be hard to understand and explain in lay terms.[68] Moreover, many Machine Learning models include multilayered networks in which the outputs are a result of an internal process that may not be replicated or understood mathematically even by the data scientists and the solution designers themselves.[69] This multilayered architecture is commonly known as the “black box”, since it may make it impossible for those using the solution to understand how it reached a specific conclusion or prediction. In other words, the reasoning behind the functioning of these applications is in most cases not transparent or intelligible for human beings; consequently, it is difficult to assess the fairness and quality of the process.

One suggested answer to the challenge of transparency in Artificial Intelligence applications is to explain the logic behind the solutions, in other words giving information about the type of input data and the expected output, explaining the variables and their weight, or shining light on the analytics architecture. This approach, known as “interpretability”, focuses on understanding the causality of a change in the input to the output, without necessarily explaining all the logic of the machine through its multiple layers. In the case of black boxes, however, achieving interpretability will often be difficult and it is important to be transparent with Data Subjects about unknowns and areas of uncertainty. Other approaches are based on selective disclosure or contractual strategies, but they also suffer some limits or cannot be generalized.[70]

Humanitarian Organizations need to work with developers on the issue of “explainability”, especially when they intend to use Artificial Intelligence solutions to support decision making. They should be able to explain to Data Subjects how the solution works, what risks may arise, how the Artificial Intelligence system achieves its outcomes and what arrangements are in place for a human decision maker to review its decisions or suggestions if needed.

Finally, care should be taken in decision making about transparency if it conflicts with data sensitivity at the individual level, or when transparency in Processing could encourage circumvention of the data Processing system by malicious actors and thus bias it.

17.2.5 DATA MINIMIZATION

The data minimization principle requires organizations to limit the Processing of Personal Data to the minimum amount and extent necessary to achieve the purpose of the Processing.[71] With the use of Artificial Intelligence, however, large-scale

[68] The Norwegian Data Protection Authority, *Artificial Intelligence and Privacy*, 19.

[69] Future of Privacy Forum, *The Privacy Expert’s Guide to Artificial Intelligence and Machine Learning*, 17.

[70] See also Andrew D. Selbst and Solon Barocas, “The intuitive appeal of explainable machines”, *Fordham Law Review*, Vol. 87, No. 3, 2019 2018, pp. 1085–1140.

[71] See Section 2.5.4 – The principle of data minimization.

Processing is often required for its functioning, and moreover the search for new patterns and correlations in data sets can make it difficult to circumscribe the range of data used.[72] Moreover, training such solutions using suitably large and representative data sets is also necessary to reduce potential bias in their outcomes.[73]

Despite this tension between Artificial Intelligence and data minimization, various solutions are possible to balance the different needs. These are set out below, along with their potential limitations:

- Employing techniques that can make it harder to identify individuals through the data, such as restricting the amount and nature of the information used. This approach may not fit certain Artificial Intelligence solutions that require large amounts of data to function well. In addition, making data hard to identify does not, by itself, guarantee respect for the data minimization principle.
- Using "synthetic data" as training data. Synthetic data "is an artificial data set, including the actual data on no 'real' individuals, but which mirrors in characteristics and proportional relationships all the statistical aspects of the original dataset".[74] This is a very promising solution,[75] but it still requires real data as a starting point. It also requires more expertise from data scientists, and it may suffer from some limitations stemming from the replication process and the difficulty of ensuring accuracy when many variables and complex situations are considered.
- Adopting a progressive approach by collecting what is thought to be the minimum amount of data necessary to achieve the expected results and then testing the solution to see how it performs. After testing, more data may be added if needed, and the solution can be tested again until it achieves the desired outcomes. This approach reduces the Processing of unnecessary data and seeks to ensure that the solution is trained on the minimum possible data set, while also making Reidentification harder.

Despite the challenges associated with data minimization in Artificial Intelligence, this principle does not mean that large-scale Processing is forbidden, but rather that it poses higher risks that require appropriate security and risk-mitigation measures. Moreover, as mentioned previously, not all Artificial Intelligence solutions require large volumes of data to be accurate. Those based on reinforcement learning, for instance, can be trained with little data.

The data processed by Humanitarian Organizations should be adequate and relevant for the purposes for which they are collected and processed. This means ensuring

[72] Centre for Information Policy Leadership, *Artificial Intelligence and Data Protection in Tension*, 14.

[73] Ibid., 13.

[74] Future of Privacy Forum, *The Privacy Expert's Guide to Artificial Intelligence and Machine Learning*, 8.

[75] Council of Europe (CoE), *Artificial Intelligence and Data Protection: Challenges and Possible Remedies | T-PD(2018)09Rev*.

that data collection is not excessive and that the time period for which the data are stored, before being anonymized or archived, is limited to the minimum necessary. The amount of Personal Data collected and processed should, ideally, be limited to what is necessary to fulfil the specified purpose(s) of data collection, data Processing or compatible Further Processing, or to what is justified on another legal basis.

Finally, although Artificial Intelligence often requires large-scale data sets, it is always crucial to carefully design the data strategy, by keeping the contents of data sets collected by Humanitarian Organizations to the minimum necessary for the purposes of the Processing and defining the purpose of data Processing as specifically as possible. Data Controllers and, where applicable, Data Processors should carefully consider the design of their data analysis, in order to minimize the presence of redundant and marginal data.[76]

17.2.6 DATA RETENTION

Personal Data should be retained only for a defined period as necessary for the purposes for which they were collected.[77] Following the initial retention period an assessment should be made as to whether the data should be deleted or whether they should be kept for a longer period to achieve the purpose. If this Processing is performed on pre-existing data sets, as “compatible Further Processing”,[78] the Processing should take place within the data retention period allowed for the purpose of initial collection. Renewal of the initial retention period, if a renewal is contemplated by the retention policy at the time of collection, can take place to enable analytics as “compatible Further Processing”.

However, in the Artificial Intelligence context, a longer period for data retention may be justified when data are used to monitor the performance system[79] and prevent unexpected biases. If a model shows bias, it can be helpful to have the training data set available to investigate the potential source of the bias. During the retention period, Data Controllers must ensure that data remain updated to reduce the risk of inaccuracies.[80]

Given the variety of uses Artificial Intelligence may have in the humanitarian sector, specific retention periods should be considered in the context of each programme. In this regard, Humanitarian Organizations should consider and set an initial retention period, such as a two-year period for audit purposes. Should the data still be

[76] Council of Europe (CoE), *Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data | T-PD(2017)01*.

[77] See Section 2.7 – Data retention.

[78] See Subsection 2.5.2.1 – Further Processing.

[79] Centre for Information Policy Leadership, *Artificial Intelligence and Data Protection in Tension*, 15.

[80] Article 29 Data Protection Working Party, *Guidelines on automated individual decision-making*, 12.

needed after this initial period, organizations should conduct periodic assessments based on their retention needs and consider their legal basis for amending the retention period. They will also need to seek additional Consent from Data Subjects if their data are retained for longer than the duration they consented to at the point of collection.

17.2.7 DATA SECURITY

Data security[81] is an essential aspect of Artificial Intelligence solutions, particularly in the humanitarian sector. Humanitarian Organizations must be mindful of the risks that these technologies pose and implement the highest level of data security when using them. Attacks by malicious parties typically fall into one of three categories:

- **model inversion attacks**: attempts to reveal information about the training data by inverting the system's model;
- **poisoning attacks**: attempts to decrease the utility of the model;
- **backdoor attacks**: attempts to gain unauthorized access to the solution and modify it after it has been trained.

Looking specifically at model inversion, it has been demonstrated that some systems remember their training data sets. For example, if a person's face has been used to train a facial recognition system, a malicious party could query the system again and again, slowly changing the input image to reconstruct the face with sufficient precision to know that the person in question was part of the training set.[82]

Another type of deliberate attack involves adding noise to the data in order to decrease the quality of outcomes, sometimes even leading to useless results such as making wrong classifications and predictions.

All these factors mean that inadequate data security can pose significant risks for vulnerable individuals in the context of the use of Artificial Intelligence. In view of these risks, it is important to build strong and secure systems that effectively protect against unauthorized access. Pseudonymization and encryption techniques are some of the methods that can assist in this regard. While the technique of training models on encrypted data is still in its early days, static models that receive encrypted inputs and produce encrypted outputs are already commonplace, albeit with their own constraints. The use of differential privacy[83] should also be considered when training Artificial Intelligence solutions.

[81] See Section 2.8 – Data security and Processing security.

[82] Matt Fredrikson, Somesh Jha and Thomas Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures", in *CCS'15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, ACM, Denver, CO, 2015, 1322–1333: doi.org/10.1145/2810103.2813677.

[83] "Differentially-private algorithms are resilient to adaptive attacks that use auxiliary information. These algorithms rely on incorporating random noise into the mix so that everything an adversary receives

Finally, in considering the suitability of security measures required to protect information in Artificial Intelligence-based solutions, it is important to take into account that the outputs of the Processing may produce more Sensitive Data than the initial data sets, including individual or group profiling, and could prove harmful to the individuals concerned if they fall into the wrong hands. In this case, the Humanitarian Organization should implement adequate security measures to protect the output, which are appropriate for the risks involved.[84] Additionally, regular data security and data privacy training is essential to raise awareness of security threats and to avoid Data Breaches.

17.3 RIGHTS OF DATA SUBJECTS

Data Controllers are responsible for determining the means and purposes of the Processing and for ensuring that Data Subjects can exercise their rights.[85] Although Artificial Intelligence may make it more difficult for Data Controllers to comply with these obligations, choosing such solutions as a means to achieve a certain purpose does not excuse Data Controllers from their responsibilities. Humanitarian Organizations should therefore have procedures and systems in place to ensure that individuals can exercise their rights. At the same time, as is discussed in Section 2.11 – Rights of Data Subjects, the exercise of these rights may be limited in certain circumstances.

17.3.1 RIGHTS RELATED TO AUTOMATED DECISION MAKING

Data Subjects have the right to not be subjected to solely automated decision making, i.e. "decisions by technological means without human involvement",[86] when such decisions produce legal effects or similarly significantly affect the individual in question.

EXAMPLE:
Some examples of solely automated decision making include speeding fines imposed purely on the basis of evidence from speed cameras, automatic refusal of an online credit application or e-recruiting practices without any human intervention.[87]

becomes noisy and imprecise, and so it is much more difficult to breach privacy (if it is feasible at all)". Aaruran Elamurugaiyan, "A Brief Introduction to Differential Privacy" Medium, 31 August 2018: https://medium.com/georgian-impact-blog/a-brief-introduction-to-differential-privacy-eacf8722283b.

[84] See Section 17.2.7 – Data security, and Section 2.8 – Data security and Processing security.

[85] See Section 2.11 – Rights of Data Subjects.

[86] Council of Europe (CoE), *Artificial Intelligence and Data Protection: Challenges and Possible Remedies | T-PD(2018)09Rev*, 8.

[87] Ibid.

The rationale behind this right "is driven by a concern for algorithmic bias; a worry of incorrect or unsubstantiated solely automated decisions based on inaccurate or incomplete data; and the need for individuals to have redress and the ability to contest a decision if an Artificial Intelligence system is incorrect or unfair".[88] These concerns are justified by examples such as the Swedish benefits case mentioned above, where a rogue solution meant that "thousands of unemployed people were wrongly denied benefits".[89] In Humanitarian Action, a similar problem could arise if Artificial Intelligence solutions make decisions about who receives aid or who is included in a target population for an aid programme. Beneficiaries should always have the right to have a human being oversee decisions that affect them.

It should be noted that "[t]o qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture".[90] This is particularly important because those making decisions may blindly rely on the Artificial Intelligence solution's suggestions on the basis that mathematical algorithms are supposedly failproof. Consequently, the presence of an individual human decision maker alone is not sufficient. The decision maker must have the ability to refute the machine's decision or suggestion.[91]

On a similar note, decision makers may not fully understand how the system arrived at a particular decision or suggestion and may therefore find it difficult to assess whether it was made wrongly (see Section 17.2.4 – Transparency, above). Decision makers should always be able to examine all the facts and information from scratch and make an independent decision, without considering the Artificial Intelligence solution's outcome. This is not always straightforward, however, since an Artificial Intelligence solution is able to process much more information than a person in the same situation. Setting up a multidisciplinary team, including individuals with expertise in the sector and technology developers, may be one option in such cases.

It is possible that individuals, regardless of their level of expertise, may be reluctant to challenge an Artificial Intelligence system's automated decisions, given how accurate the technology can be. Consequently, another issue to take into account is how the human intervention would be arranged so that a review of the decision is "carried out by someone who has the appropriate authority and capability to change the decision".[92] Organizations therefore need to consider whether it would be acceptable

[88] Centre for Information Policy Leadership, *Artificial Intelligence and Data Protection in Tension*, 16.

[89] Willis, "Rogue Algorithm Stops Welfare Payments".

[90] Article 29 Data Protection Working Party, *Guidelines on automated individual decision-making*, 21.

[91] Council of Europe (CoE), *Artificial Intelligence and Data Protection: Challenges and Possible Remedies | T-PD(2018)09Rev*.

[92] Article 29 Data Protection Working Party, *Guidelines on automated individual decision-making*, 27.

for beneficiaries to be subjected to automated decision making if they had the right to request human intervention. Here, the very case for using the technology in the first place may come under challenge.

In any case, it is essential that beneficiaries are informed about any automated decision making they are being subjected to, including the logic behind the Artificial Intelligence solution, the significance of the Processing and its envisaged consequences for them.[93] They must also be able to object to the Processing.

The rights of the Data Subjects are described in Section 2.11 – Rights of Data Subjects. The rights to information, access, correction, erasure and objection are considered crucial components of an effective data protection policy. However, Artificial Intelligence-based Processing of Personal Data poses significant challenges.

The Data Subject's exercise of the right to information about automated decision making (also relevant to the transparency principle, see Section 17.2.2 – Purpose limitation and Further Processing) is more difficult in the Artificial Intelligence context, given the complexity of such systems and how they operate. It is therefore important to explore alternative means of Artificial Intelligence transparency and consider new forms of information provision, such as the creation of public registers describing the key functions and characteristics of the most impactful systems. It may also be advisable to investigate the provision of information to representatives of potentially affected groups.

Organizations engaged in humanitarian use of Artificial Intelligence are encouraged to incorporate complaint procedures into their Personal Data Processing practices and internal data protection policies. These procedures should enable data correction and erasure. However, it should be recognized that the exercise of certain individual rights may be limited by the legal basis of the Processing. For example, requests for opt-outs by individuals may not be observed in the event of Processing undertaken under the legal basis of public interest described above.

Humanitarian Organizations need to ensure that no automated decisions are taken with regard to individuals which could lead to harm or exclusion from humanitarian programmes, without any human intervention. In practice, this means that a human being should always be the final decision maker when decisions are taken on the basis of Artificial Intelligence outputs that may have adverse effects on individuals.

[93] Ibid., 25.

EXAMPLE
In the event of aid distribution, a decision based on output from Artificial Intelligence to prioritize a specific region or group of people (to the disadvantage of those left out of these regions or groups) should always be cross-checked and validated by a human being.

17.4 DATA CONTROLLER/DATA PROCESSOR RELATIONSHIP

Artificial Intelligence solutions tend to blur the traditional distinction between the roles of Data Controller and Data Processor, which is centred on the idea of power to control and supervise the data Processing in relation to the definition of its purposes and means. This is largely due to the fact that in the case of Artificial Intelligence solutions, providers retain important privileges as regards the organization of the service and Artificial Intelligence architecture.

17.4.1 ACCOUNTABILITY

To have a proper allocation of accountability and liability obligations, it is crucial to carefully determine which entity actually acts as Data Controller, retaining the control over personal information and a general power to manage the purposes and means of data Processing, and which processes Personal Data on behalf of the Data Controller and is therefore a Data Processor. It is also possible that more than one entity jointly determines the purposes and means of the Processing and may be considered as joint Data Controllers.

EXAMPLE 1: Humanitarian Organizations sharing data sets and undertaking Data Analytics using their own organizational resources may be considered joint Data Controllers.
EXAMPLE 2: Humanitarian Organizations sharing data sets but outsourcing the Data Analytics to a commercial service provider that will transfer the findings and keep no records for its own use will be considered joint Data Controllers, and the service provider will be considered a Data Processor.

In accordance with their different roles and respective spheres of competence, Data Controller and Data Processor are accountable for the decisions they adopt concerning data Processing. However, as explained above, Artificial Intelligence sometimes evolves in ways that cannot be fully understood by developers themselves due to the “black box” effect. This may raise questions around the concrete implementation of the accountability principle, which requires Data Controllers to comply with data protection requirements and to be in a position to demonstrate that they have taken adequate and proportionate technical and organizational measures within their respective Processing operations.[94]

17.4.2 LIABILITY

Automated decision making (see above) raises particular issues around liability. In health care, for instance, machines are often considered to be more accurate than humans in diagnosing certain diseases such as specific types of cancer, or at analysing X-ray images. For this reason, doctors may feel compelled to follow the machine's recommendation.[95] Here, it might be unclear who is responsible for the diagnosis.[96] To counterbalance this, organizations may seek to extend the product liability logic to algorithms, thereby placing the full burden of liability on the developer company (although this may be very difficult to negotiate in practice). From an ethical perspective, it is also important for Humanitarian Organizations to understand their own responsibilities when choosing to use such technology and to be accountable to beneficiaries accordingly.

In a different scenario, the performance of Artificial Intelligence systems can be significantly affected by the poor quality of data available in a given context, such as in geographic areas where the use of poor scanning technologies generates biases in image-based diagnoses. In these cases, Humanitarian Organizations must therefore carefully assess the data quality to avoid potential liability.

Some specific tools, such as a data management plan and DPIA, can contribute to better clarify the roles of different parties engaged in the Processing. Once these roles have been defined and the corresponding tasks assigned, it is important to establish which relevant contracts need to be entered into among the data Processing participants.

Data collection or International Data Sharing across Humanitarian Organizations and/or national borders and/or third (private or state) bodies should generally be covered by contractual clauses. These contracts are important and can play a key role in liability management for the following reasons:

[94] See Section 2.9 – The principle of accountability.

[95] Victor Demiaux and Yacine Si Abdallah, "Comment permettre à l'homme de garder la main? Les enjeux éthiques des algorithmes et de l'intelligence artificielle", French Data Protection Authority (CNIL), Paris, December 2017, 27: www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf.

[96] Ibid.

- They should clearly allocate the roles between the various parties and, in particular, put them on notice as to whether they are acting as Data Controllers, Data Processors or joint controllers.
- They should contain an outline of the data protection obligations to which each party is subject. This should include the measures that the parties should take to protect Personal Data transferred across borders.
- They should contain obligations to cover data security, responses (objection or notification to the other party) in case of authorities requesting access to data, procedures for handling Data Breaches, Data Processor return/disposal of data at the end of the Processing, and staff training.
- They should also require that notice be given to the Humanitarian Organizations involved if any data are accessed without authorization.

17.5 INTERNATIONAL DATA SHARING

Personal Data and other types of data processed in Artificial Intelligence solutions often cross national borders due to the presence of international service providers and the use of cloud computing services. This leads to the application of provisions and practices relating to international cross-border data flows.[97] In this regard, attention must be paid to applicable law and jurisdiction.

International data sharing may involve several scenarios:

- Personal Data are transferred by a Humanitarian Organization (Data Controller) to Third Parties (Data Processors), either commercial entities or other Humanitarian Organizations, to be processed in its behalf, e.g. cloud computing service providers;
- Personal Data are shared among Humanitarian Organizations, public authorities and/or commercial entities (joint Data Controllers), e.g. partnership in joint actions;
- Personal Data are transferred to other Humanitarian Organizations, public authorities and/or commercial entities that autonomously process such information for their own purposes (Data Controllers).

Data protection laws restrict International Data Sharing, so Humanitarian Organizations should have mechanisms in place to provide a legal basis for it when Data Analytics are conducted, as discussed above.[98] It is essential to assess the potential data transfer risks prior to International Data Sharing, taking into account the local regulations in the country of destination, and to inform Data Subjects adequately. In case of potential risks, suitable mitigating measures can be adopted, both at contractual level

[97] See Chapter 4: International Data Sharing.

[98] See Section 17.2.1 – Legal bases for Personal Data Processing.

(e.g. contractual clauses, codes of conduct) and at technical level (e.g. data encryption, strong Pseudonymization). When the risk is high and the mitigation measures cannot reduce it, a decision should be taken to refrain from data sharing.[99]

Since in many cases International Data Sharing concerns the use of Third Party services, when Humanitarian Organizations hire Artificial Intelligence service providers, they should collect all relevant information on cross-border data transfers. In some cases, companies providing Artificial Intelligence solutions may have an incentive to use and exploit the results of the Processing of Humanitarian Organizations' data (e.g. commercial purposes, profiling). It is therefore very important that any contractual arrangements with them make it completely clear that the purpose of the Processing is and must remain exclusively humanitarian, and that the service provider keeps the humanitarian Processing segregated from its commercial activities.

If any doubts arise as to whether the service provider can or will respect this condition, the Humanitarian Organization should refrain from engaging in the Processing. This is because any Processing other than Processing exclusively for Humanitarian Action may have serious implications for Data Subjects. For example, outputs of analytics which identify categories of potential beneficiaries of Humanitarian Action may lead to consequences such as denial of credit, higher insurance premiums, stigmatization, discrimination or even persecution.

Humanitarian Organizations should also be alert to the risk that, in situations of violence or conflict, the parties involved may seek to access and use the findings of Artificial Intelligence-based analyses to gain an advantage, which would compromise the safety of the Data Subjects and the neutrality of Humanitarian Action. Consequently, in cases where the outputs are potentially sensitive, it is important to consider a scenario where Humanitarian Organizations develop their own Artificial Intelligence applications without recourse to Third Party solutions.

17.6 DATA PROTECTION IMPACT ASSESSMENT AND HUMAN RIGHTS IMPACT ASSESSMENT

Since the use of Artificial Intelligence can pose substantial data protection risks to individuals, an organization should carry out a Data Protection Impact Assessment (DPIA) before making a decision to implement such a solution.

A DPIA involves identifying, evaluating and addressing the impacts on Data Subjects and their Personal Data of a project, policy, programme or other initiative that entails

[99] See Chapter 4: International Data Sharing, and Section 4.4 – Mitigating the risks to the individual.

the Processing of such data.[100] It should ultimately lead to measures that avoid, minimize, transfer or share risks associated with the Processing activities. A DPIA is a continuous process and should follow a project or initiative that involves the Processing of individuals' data throughout its life cycle.

Given the limits to transparency in the use of Artificial Intelligence, publicly available DPIAs can also help increase beneficiaries' acceptance and use of Artificial Intelligence solutions by Humanitarian Organizations.

DPIAs are important tools during project design to ensure that all aspects of applicable data protection regulations and potential risks are covered.[101] DPIAs are now required in many jurisdictions and by some Humanitarian Organizations.

Apart from clarifying the details and specifications of the Processing, DPIAs should focus on the risks posed by it and on mitigating measures. These risks, according to the most relevant models of DPIA, are not limited to the right to privacy and data protection but should include risks to the rights and freedoms of natural persons.[102] In line with the by-design approach and the minimization of data Processing-related risks, DPIAs need to be conducted prior to any Artificial Intelligence-based operations and updated when Processing operation or contextual elements change.

Several risks can be considered in a DPIA including, according to the specific Processing operations, the nature of processed data, the inferences extracted using Artificial Intelligence applications, and the context where Processing is carried out. Some examples concern the risk of Reidentification of individuals of relevance for Humanitarian Action, in case of use of anonymized data or pseudonymized/aggregate results made available to Third Parties, or the risk that the results of Artificial Intelligence-based analysis performed by Humanitarian Organizations may be exploited by commercial Third Parties and/or authorities for unrelated purposes.

Further examples of risks that should be considered in the broader context of human rights protection include:

- requests to Humanitarian Organizations for specific patterns or information about certain categories of individuals by authorities or corporations that could potentially expose Data Subjects to discrimination or detrimental consequences and compromise the neutrality of Humanitarian Action;

[100] See Chapter 5: Data Protection Impact Assessments (DPIAs).

[101] See Chapter 5: Data Protection Impact Assessments (DPIAs).

[102] For assessment tools specifically developed to assess the risks in Humanitarian Action, see UN Global Pulse, *Tools: Risks, Harms and Benefits Assessment*.

- access and use of the results of Artificial Intelligence-based analysis by parties in a situation of violence or conflict to gain an advantage over other stakeholders and thus compromise the safety of the Data Subjects and the neutrality of Humanitarian Action.

Finally, considering the role of Artificial Intelligence service providers in Humanitarian Action, the DPIA should also consider the risk that commercial providers may have incentives to exploit the findings of the Processing for commercial purposes, e.g. to improve their understanding of their current or potential customers or for further customer profiling.[103]

With regard to the risk identified in the DPIA, the assessment considers the likelihood and severity of potential negative impacts on Data Subjects, also considering competing rights and freedoms and legitimate interests recognized by the law. On the basis of the analysis of this potential impact, specific mitigation measures are adopted, including in the design of the used solutions, such as Anonymization techniques, privacy-enhancing technical measures, and legal and contractual obligations to prevent possible Reidentification of the persons concerned.[104]

Although DPIA has become a mandatory requirement under national and international[105] law, assessment methodologies mainly adopt a limited perspective with a main focus on Processing, task allocation, data quality and data security, without adequately considering all the human rights potentially impacted by Artificial Intelligence applications, their diversity and complexity. However, as pointed out by the UN High Commissioner for Human Rights,[106] it is necessary to adopt a broader perspective, embedding human rights in Artificial Intelligence development, deployment and use, with a comprehensive by-design approach to counter potential adverse impacts.

17.6.1 HUMAN RIGHTS IMPACT ASSESSMENT FOR ARTIFICIAL INTELLIGENCE

Human Rights Impact Assessment (HRIA) can thus guide Artificial Intelligence developers and users from the outset in the design of new Artificial Intelligence solutions, facilitating comparison between alternative design options, and following the product/service throughout its life cycle, by using an iterative approach, based on risk assessment

[103] See Section 2.3: Aggregate, Pseudonymized and Anonymized data sets.

[104] Council of Europe (CoE), *Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data | T-PD(2017)01*.

[105] Council of Europe (CoE), Convention 108; Council of Europe (CoE), Protocol Amending the Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, para. 10.

[106] Office of the United Nations High Commissioner for Human Rights (OHCHR), *A/HRC/48/31: The Right to Privacy in the Digital Age*, report of the United Nations High Commissioner for Human Rights, UN Doc, OHCHR, 15 September 2021, www.ohchr.org/en/documents/thematic-reports/ahrc4831-right-privacy-digital-age-report-united-nations-high.

and design mitigation solutions. For these reasons, HRIA is considered the cornerstone of future Artificial Intelligence regulation at international and regional level.[107]

However, in dealing with the impact of Artificial Intelligence, traditional HRIA methodologies cannot be applied directly but must be contextualized by considering the specific nature of Artificial Intelligence. The two most relevant changes introduced in the HRIA in relation to the Artificial Intelligence context concern the *ex ante* nature of the assessment carried out and the greater focus on quantifiable risk thresholds. As for the former, an *ex ante* approach is required by the guiding role that HRIA aims to play in Artificial Intelligence project design and development, as opposed to the *ex post* evaluation centred on corrective policies that usually characterizes traditional HRIA.

Regarding the focus on risk thresholds, this is in line with the requirements emerging in the regulatory debate on Artificial Intelligence where the definition of different risk levels is crucial in the acceptability of Artificial Intelligence products and services, and directly impacts on the obligations of Artificial Intelligence manufacturers, providers and users. A quantitative dimension of assessment, in terms of ranges of risks, is therefore needed both for Artificial Intelligence design guidance and legal compliance.

Compared to the voluntary and policy-based traditional HRIA practice in the business sector, once HRIA becomes a legal tool it is no longer merely a source of recommendations for better business policy. Future Artificial Intelligence regulation will most likely bring specific legal obligations and sanctions for non-compliance in relation to risk assessment and management, as well as given risk thresholds (e.g. high risk).

17.6.2 HUMAN RIGHTS IMPACT ASSESSMENT: PHASES AND PROCEDURE

Notwithstanding these important differences impacting on the assessment methodology, the main building blocks of the assessment procedure remain the same and are similar to the phases of DPIA schemes: (i) the planning and scoping phase and (ii) the data collection and analysis phase.

The first stage deals with the definition of the HRIA target, identifying the main features of the product/service and the context in which it will be placed. There are three main areas to consider at this stage: (i) description and analysis of the type of product/service; (ii) analysis of the human rights context; (iii) identification of relevant stakeholders.

[107] CoE (Ad Hoc Committee on Artificial Intelligence (CAHAI)) "*5th Meeting, Strasbourg, 5–7 July 2021: Abridged Meeting Report and List of Decisions*", 7 July 2021, CAHAI(2021)10: www.coe.int/en/web/artificial-intelligence/cahai-1; European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, 21 April 2021, COM/2021 206 final.

The second stage focuses on relevant empirical evidence to assess the impact on human rights. Since in most cases the assessment is not based on measurable variables, the impact on rights and freedoms is necessarily the result of expert evaluation, where expert opinion relies on knowledge of case law, the literature and the legal framework. This means that it is not possible to provide a precise measurement of the expected impacts but only an assessment in terms of range of risk.

In line with risk assessment procedures, three key factors must be considered: risk identification, likelihood (L) and severity (S). With regard to the first, the focus on human rights and freedoms already defines the potentially affected categories and the case-specific analysis identifies those concretely affected, depending on the technologies used and their purposes. Since this is a rights-based model, risk concerns the prejudice to rights and freedoms, in terms of unlawful limitations and restrictions, regardless of material damage.

The expected impact of the identified risks is assessed by considering both the likelihood and the severity of the expected consequences, using a four-step scale (low, medium, high, very high) to avoid any risk of average positioning.

Likelihood is the combination of the probability of adverse consequences and the exposure (Table 17.3). The former concerns the probability that adverse consequences of a certain risk might occur (Table 17.1) and the latter the potential number of people at risk (Table 17.2). Both these variables must be assessed on a contextual basis and the engagement of relevant shareholders can be of help.

The severity of the expected consequences (Table 17.6) is estimated by considering the gravity of the prejudice in the exercise of rights and freedoms (Table 17.4) and the effort to overcome it and to reverse adverse effects (Table 17.5).

Table 17.1 Probability

Low	The risk of prejudice is improbable or highly improbable.	1
Medium	The risk may occur.	2
High	There is a high probability that the risk occurs.	3
Very high	The risk is highly likely to occur.	4

Table 17.2 Exposure

Low	Few or very few of the identified population of rights-holders are potentially affected.	1
Medium	Some of the identified populations are potentially affected.	2
High	The majority of the identified population is potentially affected.	3
Very high	Almost the entire identified population is potentially affected.	4

Table 17.3 Likelihood table (L)

		Probability			
		1	2	3	4
Exposure	1	1	2	3	4
	2	2	3	5	9
	3	3	5	9	12
	4	4	7	12	15

Likelihood	
Low	1
Medium	2
High	3
Very high	4

Table 17.4 Gravity of the prejudice

	Gravity of the prejudice	
Low	Affected individuals and groups may encounter only minor prejudices in the exercise of their rights and freedoms.	1
Medium	Affected individuals and groups may encounter significant prejudices.	2
High	Affected individuals and groups may encounter serious prejudices.	3
Very high	Affected individuals and groups may encounter serious or even irreversible prejudices.	4

Table 17.5 Effort to overcome the prejudice and to reverse adverse effects

	Effort	
Low	Suffered prejudice can be overcome without any problem (e.g. time spent amending information, annoyances, irritations, etc.).	1
Medium	Suffered prejudice can be overcome despite a few difficulties (e.g. extra costs, fear, lack of understanding, stress, minor physical ailments, etc.).	2
High	Suffered prejudice can be overcome albeit with serious difficulties (e.g. economic loss, property damage, worsening of health, etc.).	3
Very high	Suffered prejudice may not be overcome (e.g. long-term psychological or physical ailments, death, etc.).	4

Taking into consideration the L and S values, the overall impact is determined using a table (Table 17.7) where colours from lightest to darkest represent the overall impact, from lowest to highest. Once the potentially adverse impact has been assessed for each of the rights and freedoms considered, a radial graph (Graph 17.1) of the overall

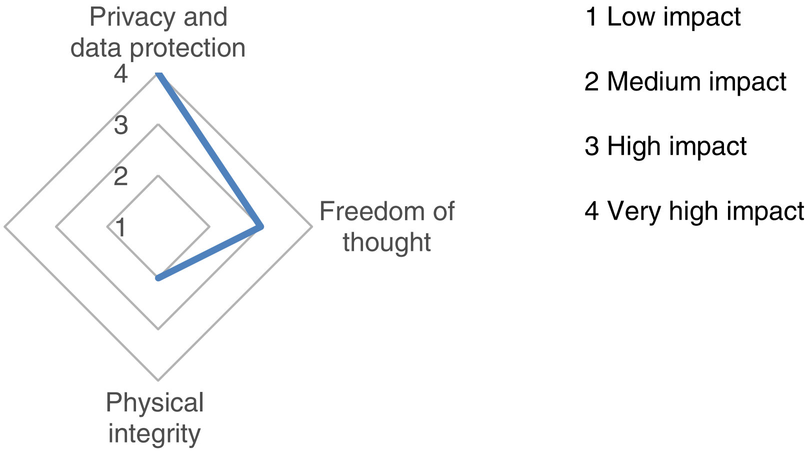
Table 17.6 Severity table (S)

		Gravity			
		1	2	3	4
Effort	1	1	2	4	6
	2	2	3	5	8
	3	3	5	8	10
	4	5	8	10	12

Severity	
Low	1
Medium	2
High	3
Very high	4

Table 17.7 Overall risk impact table

		Severity [impacted right/freedom]			
		Low	**Medium**	**High**	**Very high**
Likelihood	Low	□	░	▒	▒
	Medium	□	░	▒	▒
	High	░	▒	▒	▓
	Very high	░	▒	▓	▓



Graph 17.1. Radial graph (impact) example

impact can be used to decide the priority of intervention in altering the characteristics of the product/service to reduce the expected adverse impacts. Factors that can exclude the risk from a legal perspective (e.g. the mandatory nature of certain impacting features) should be considered.

After the first adoption of the appropriate mitigation measures for the foreseen risks, further rounds of assessment can be conducted according to the level of residual risk and its acceptability.

17.7 DATA PROTECTION BY DESIGN AND BY DEFAULT

Data Protection by design and by default involves designing a Processing operation, programme or solution in a way that implements key data protection principles from the outset, and that provides the Data Subject with the greatest possible data protections (see Chapter 6: Designing for data protection). The key data protection principles in this sense are:

- lawfulness, fairness and transparency;
- purpose limitation;
- data minimization;
- accuracy;
- storage limitation (limited retention);
- integrity and confidentiality (security);
- accountability.

The by-design approach also represents the concrete implementation of the impact assessment concerning data Processing. The adoption of specific mitigation measures or changes to the system design are usually the main way to tackle the potential risks identified in the impact assessment.

The measures to be adopted from a data protection by design perspective are necessarily context-specific, but solutions such as synthetic data, Pseudonymization, Anonymization (where possible) and encryption techniques are frequently components of the by-design approach.

17.8 ETHICAL ISSUES AND CHALLENGES

Given the speed at which technologies are evolving, the law often lags behind major societal changes. It is therefore likely that some of the ethical issues associated with Artificial Intelligence solutions are not yet covered by existing laws. In addition, there is a sphere of social and ethical issues and values that is not reflected in legal provisions but is relevant in defining a given community’s approach to the use of data-intensive Artificial Intelligence systems and their social acceptability.

When opting to develop or use Artificial Intelligence solutions, Humanitarian Organizations should of course consider whether they comply with data protection laws and data protection by design principles. Importantly, however, they should also reflect on potential adverse impacts on the ethical and social implications of the data Processing.[108] For more guidance on the topic of analysing systems, see Section 6.3.3 – Analysing purpose limitation.

Artificial Intelligence tools present many risks, such as the possibility of discriminatory bias or lack of system accuracy. Also, some developers may train systems on data obtained either illegally or through unethical methods. This is particularly worrisome when users of such platforms or services are members of vulnerable groups.

Risk assessments that go beyond traditional data protection and cover a wider range of interests, ethical standards and rights (such as the right to non-discrimination)[109] are of great importance. Societal interests and ethics are broader than law, and organizations should consider the wider contextual background, including political and cultural nuances. This makes evaluating ethical values more complex, context-dependent and comprehensive than assessing compliance with data protection laws alone.

There have been numerous attempts to define the ethical principles that apply to the development of Artificial Intelligence. Examples include the Asilomar Artificial Intelligence Principles[110] and the International Conference of Data Protection and Privacy Commissioners' *Declaration on Ethics and Data Protection in Artificial Intelligence*.[111] Academics are also conducting research into ethical issues related to Artificial Intelligence,[112] and some multinational companies are developing their own sets of ethical principles.[113]

[108] Mantelero, *Beyond Data Human Rights, Ethical and Social Impact Assessment in AI*.

[109] Ibid., chap. 2.

[110] Future of Life Institute, "Asilomar AI Principles": https://futureoflife.org/ai-principles.

[111] International Conference on Data Protection and Privacy Commissioners, *Declaration on Ethics and Data Protection in Artificial Intelligence*, Declaration, 40th International Conference of Data Protection and Privacy Commissioners, Brussels, Belgium, 23 October 2018: http://globalprivacyassembly.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf.

[112] See for example the ACM conference on Fairness, Accountability and Transparency (fatconference.org), which has gained prominence in recent years.

[113] Marcello Ienca and Effy Vayena, "AI ethics guidelines: European and global perspectives", in *Towards Regulation of AI Systems: Global Perspectives on the Development of a Legal Framework on Artificial Intelligence Systems Based on the Council of Europe's Standards on Human Rights, Democracy and the Rule of Law*, by Isaac Ben Israel et al., Council of Europe (CoE), 2020, 38–60: https://edoc.coe.int/en/artificial-intelligence/9656-towards-regulation-of-ai-systems.html; Thilo Hagendorff, "The ethics of AI ethics: An evaluation of guidelines", *Minds and Machines*, Vol. 30, No. 1, 1 March 2020, pp. 99–120: https://doi.org/10.1007/s11023–020-09517-8.

However, ethical assessment, like social assessment, is more complicated than that of Data Protection and Human Rights Impact Assessment. Whereas the latter refer to a well-defined benchmark, the ethical framework involves a variety of theoretical inputs on the underlying values, as well as a proliferation of guidelines, in some cases partially affected by "ethics washing" or reflecting corporate values.

Experts therefore play a crucial role in detecting, contextualizing and evaluating Artificial Intelligence solutions against existing ethical and social values. Much more than in the human rights assessment, experts are decisive in grasping the relevant community values, given their context-specific nature and, in many cases, the need for active interaction with rights-holders and stakeholders to better understand them.

Given the impact Artificial Intelligence can have, ethics committees are attracting increasing attention in Artificial Intelligence circles as they can provide valuable support to developers in designing rights-based and socially oriented algorithms.[114] In terms of the composition of such committees, where societal issues are significant, legal, ethical or sociological expertise, as well as domain-specific knowledge, will be essential. Humanitarian Organizations could therefore consider establishing an ethics committee to assist them in dealing with such issues when deploying Artificial Intelligence solutions.

To ensure compliance with legal and ethical standards, Humanitarian Organizations should consider the following two steps:

- First, they should answer the following three questions:
  1. What should actually be done?
  2. What is legally allowed?
  3. What is technically possible?
- Second, when choosing to use new technologies, they should consider the problem they are facing and whether Artificial Intelligence can help solve it by asking the questions below:
  - What problem is solved with Artificial Intelligence?
  - What problem is not solved?
  - What problem is created?
  - How does this technology perform compared with other technologies that may be less risky?

In this respect, ethical assessment also has an influence on the design of Artificial Intelligence solutions, especially with regard to acceptability of the proposed Artificial Intelligence solution. This assessment not only examines the Artificial Intelligence product/service itself but looks at a wider range of alternative

[114] Council of Europe (CoE), *Artificial Intelligence and Data Protection: Challenges and Possible Remedies | T-PD(2018)09Rev*, 16.

possibilities to meet identified needs, also considering solutions that are not necessarily based on Artificial Intelligence.

In this regard, the zero option (not using Artificial Intelligence) should always be kept in mind. This is particularly relevant where the use of Artificial Intelligence would be legal but not ethically acceptable. For instance, if the solution chosen by the organization is not well accepted by the intended beneficiaries of the programme, this feeling of discomfort or distrust may justify a decision not to implement the technology.