PANELFIT

PARTICIPATORY APPROACHES TO A NEW ETHICAL AND LEGAL FRAMEWORK FOR ICT

Guidelines on Data Protection Ethical and Legal Issues in ICT Research and Innovation.

ARTIFICIAL INTELLIGENCE (AI)

AI: Requirements for developers and innovators

Iñigo de Miguel Beriain[1] (UPV/EHU), Felix Schaber[2] (OEAW), Gianclaudio Malgieri and Andrés Chomczyk Penedo[3] (VUB)

Acknowledgements: The authors thankfully acknowledge José Antonio Castillo Parrilla, Eduard Fosch Villaronga and Lorena Perez Campillo for their kind support in writing this document. Needless to say, all mistakes are our full responsibility.

This part of The Guidelines has been reviewed and validated by Marko Sijan, Senior Advisor Specialist, (HR DPA)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 788039. This document reflects only the author's view and the Agency is not responsible for any use that may be made of the information it contains.

[1] Author of the whole document except sections 2 and 7 of this part.

[2] Author of section 2 of this part.

[3] Authors of section 7 of this part.

Introduction part A

The first part of this chapter is built around the seven ethical requirements included in the recommendations published by the High-Level Expert Group on AI[4] in their 'Ethics guidelines for trustworthy AI'.[5] These recommendations were recently analyzed by the SHERPA project,[6] which included an extensive analysis of the ethical issues involved when developing adequate tools to face these challenges. In light of this commendable effort, it would be redundant to include an extensive analysis here of the same topic (AI) from the same perspective (ethics). Instead, what we have tried to do in this document is provide a complementary analysis. These Guidelines aim to find the overlap between the ethical recommendations made by the High-Level Expert Group on AI and the legal framework created by the General Data Protection Regulation (GDPR) on data protection issues.

Before starting the analysis, however, it is necessary to include some preliminary notes. First, this report focuses mainly on AI developers: organizations willing to develop AI tools. These organizations become controllers as soon as they start processing personal data. In a similar vein, the terms 'tool', 'solution', 'model' and 'development' should be considered as synonymous in the context of this analysis.

Second, this part of the Guidelines can only be understood in the context of the whole tool (the Guidelines). There are several concepts that are not explored in this document, because they are addressed in other sections of The Guidelines; we have referred to these wherever needed (references are highlighted in yellow). In the future, all sections will be available on a website, making the Guidelines much more user friendly.

The different sections in this part of the document contain only what we consider to be strictly necessary to understand the core arguments of the ethical and legal issues at stake. We have included checklists that should help controllers to determine whether they are addressing the issues accurately, and a further reading section for readers to consult if necessary. Footnotes provide further references to the most important statements.

Finally, this document has been structured so that it is easy to understand. As previously mentioned, it is based upon the seven requirements described by the High-Level Expert Group on AI. We start our analysis with a brief description of the core ethical issues at

---

[4] The High-Level Expert Group on AI was created by the European Commission in 2018. It has as a general objective to support the implementation of the European Strategy on Artificial Intelligence. This includes the elaboration of recommendations on future-related policy development and on ethical, legal and societal issues related to AI, including socio-economic challenges. Available at: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence (accessed 15 May 2020).

[5] High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI, p.15 and ff. Brussels, European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 15 May 2020).

[6] SHERPA (2019) Guidelines for the ethical use of AI and big data systems. SHERPA, www.project-sherpa.eu/wp-content/uploads/2019/12/use-final.pdf (accessed 5 May 2020).

stake, and then summarize the main ethical issues and challenges related to them. These serve as the common basis on which our legal analysis is built, providing the context for the legal analysis made.

1 Human agency and oversight

"AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights and allow for human oversight."

- *High-Level Expert Group on AI*[7]

1.1 Ethical principles

This first requirement for the development of AI embeds three main different principles:[8]

- **Fundamental rights.** AI systems can enable or hamper fundamental rights. Under such circumstances, a fundamental rights impact assessment should be undertaken before developing an AI solution.
- **Human agency.** Users of AI systems should be able to make informed, autonomous decisions about doing so. AI systems should support individuals in making better, more informed choices in accordance with their goals. The overall principle of user autonomy must be central to the AI system's functionality. For example, data subjects must be aware that their data could be used for profiling, if this might happen. Furthermore, their right not to be subject to a decision based solely on automated processing, when this produces legal effects or similarly significantly effects, must be respected. However, one must keep in mind that this, in general, refers to commercial purposes. Thus, the same is not applicable as for Law Enforcement Agencies that process personal data on legal basis and might use AI for effective fight against different crimes and to fulfil obligations stipulated by law
- **Human oversight.** Human oversight helps to ensure that an AI system does not undermine human autonomy or cause other adverse effects. Such oversight may be achieved through diverse governance mechanisms. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required.

[7] High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI, p.15 and ff. Brussels, European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 15 May 2020).

[8] Ibid., p.15.

1.2 GDPR provisions

The requirement for human agency and oversight when developing AI tools is clearly linked to the right to obtain human intervention, the right not to be subject to a decision based solely on automated processing, the right to information on automatic decision-making (ADM), and the logic involved, which are all included in the GDPR. These rights are challenged by the use of AI tools. AI often involves a form of automated processing and, in some cases, decisions are directly taken by the AI model. Indeed, sometimes AI tools learn and make decisions without human supervision and sometimes the logic involved in their performance is hard to understand.[9]

In this respect, **profiling is particularly problematic in AI development** (see Box 1), because the process of profiling "is often invisible to the data subject. It works by creating derived or inferred data about individuals – 'new' personal data that has not been provided directly by the data subjects themselves. People have vastly different levels of comprehension of this subject and may find it challenging to understand the sophisticated techniques involved in profiling and automated decision-making processes"[10] (see section 'y

Understanding transparency and opacity').

Of course, the **GDPR does not prevent any form of profiling and/or automated decision-making**: it only provides individuals with a qualified right to be informed about it, and a right not to be subject to a decision based on purely automated decision-making, including profiling. Their right to information (see "Right to information" within Part II section "Data subjects' rights" of these Guidelines) must be satisfied through application of the lawfulness, fairness and transparency principle (see "Lawfulness, fairness and transparency principle" within Part II section "Principles" of these Guidelines). This means that, **as a minimum,** controllers have to inform the data subject that "they are engaging in this type of activity, provide meaningful information about the logic involved and the significance and envisaged consequences of the profiling for the data subject"[11] (see Articles 13 and 14 of the GDPR).

The **information about the logic of a system,** and explanations of decisions, should give individuals the necessary context to make decisions about the processing of their personal data. In some cases, insufficient explanations may prompt individuals to resort to other rights unnecessarily. Requests for intervention, the expression of views, or objections to the processing are more likely to happen if individuals do not feel they have a sufficient understanding of how the decision was reached. Data subjects must be able to **exercise their rights in a simple and user-friendly manner.** For example, "if the result of a solely automated decision is communicated through a website, the page should contain a link or clear information allowing the individual to contact a member

[9] Burrell, J. (2016) 'How the machine 'thinks': understanding opacity in machine learning algorithms', *Big Data & Society* 3(1): 1-12.

[10] Article 29 Working Part (2017) Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679, WP 251, p.9. European Commission, Brussels.

[11] Ibid., pp.13-14.

of staff who can intervene, without any undue delays or complications".[12] The full scope of the information to be provided is hard to state concretely, however. Indeed, there is a lively academic discussion about this issue at present.[13]

Box 0. The issue of ranking

Services or goods providers that participate in the so-called 'collaborative economy' (or 'platform economy') need to understand the functioning of ranking in the context of their use of specific online intermediation services or online search engines. This could be, for example, a hotel – whether big or small – offering its accommodation through Booking.com or TripAdvisor. To allow businesses to participate as providers on the platform, it is not necessary for platforms to disclose the detailed functioning of their ranking mechanisms, including the algorithms used. It is sufficient to provide a general description of the main ranking parameters (including the possibility to influence ranking against any direct or indirect remuneration paid by the provider), as long as this description is easily and publicly available, and written in clear and intelligible language.[14]

Furthermore, according to article 22(1), data subjects shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning them or similar significantly affects them. Thus, controllers should always make sure that the AI tools they use or develop **are not promoting unavoidable automatic decision-making in any way.** Indeed, according to the Article 29 Working Party, "[i]f the controller envisages a 'model' where it takes solely automated decisions having a high impact on individuals based on profiles made about them and it cannot rely on the individual's consent, on a contract with the individual or on a law authorising this, the controller should not proceed. The controller can still envisage a 'model' of decision-making based on profiling, by significantly increasing the level of human intervention so that the model is no longer a fully

[12] ICO (2020) AI auditing framework: draft guidance for consultation, p.94. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (accessed 15 May 2020).

[13] Wachter, S., Mittelstadt, B. and Floridi, L. (2017) 'Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation', *International Data Privacy Law*. Available at https://ssrn.com/abstract=2903469 or http://dx.doi.org/10.2139/ssrn.2903469 (accessed 15 May 2020); Selbst, A.D. and Powles, J. (2017) 'Meaningful information and the right to explanation', *International Data Privacy Law* 7(4): 233-242, https://doi.org/10.1093/idpl/ipx022 (accessed 15 May 2020).

[14] EU Regulation 1159/2019 of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services, Article 5 and Recital 27. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R1150&from=EN

automated decision-making process, although the processing could still present risks to individuals' fundamental rights and freedoms."[15]

Box 1. Understanding profiling

Research by Kosinski et al. (2013)[16] showed that, in 2011, accessible digital records of behavior (such as pages 'liked' on Facebook) could be used to accurately predict a range of highly sensitive personal attributes. These included: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age and gender. The analysis was based on a dataset of over 58,000 volunteers who provided their Facebook 'likes', detailed demographic profiles, and the results of several psychometric tests.

The model correctly discriminated between homosexual and heterosexual men in 88% of cases; between African Americans and Caucasian Americans in 95% of cases; and between Democrat and Republican voters in 85% of cases. For the personality trait, 'Openness', the prediction accuracy was close to the test–retest accuracy of a standard personality test. The authors also provided examples of association between attributes and 'likes' and discussed implications for online personalization and privacy.

This case constitutes and excellent example of how profiling works: data subjects' information served well to classify them and make predictions about them.

Furthermore, a controller must always remember that, according to Article 9(2)(a) of the GDPR, automated decisions that involve processing special categories of personal data are permitted only if the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, or if there is legal basis for the mentioned processing. This exception applies not only when the observed data fit into this category, but also if the alignment of different types of personal data can reveal sensitive information about individuals, or if inferred data fall into that category. Indeed, **a study able to infer special categories of data is subject to the same legal obligations, pursuant to the GDPR, as one in which sensitive personal data are processed from the outset.** In all such cases, we must consider the regulations applying to the processing of special categories of personal data and the necessary application of appropriate safeguards, able to protect the data subjects' rights, interests and freedoms. Proportionality between the aim of research and the use of special categories of data must be guaranteed. Furthermore, controllers must be aware that their Member States may maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health (Art 9 (4) GDPR).

---

[15] Article 29 Working Party (2018) Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679. European Commission, Brussels, p. 30. Available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053.

[16] Kosinski M., Stillwell, D. and Graepel, T. (2013) 'Digital records of behavior expose personal traits', *Proceedings of the National Academy of Sciences* 110(15): 5802-5805, DOI:10.1073/pnas.1218772110

.

If profiling infers personal data that were not provided by the data subject, the controllers must ensure that the processing is not incompatible with the original purpose (see "Data protection and scientific research" in Part II section "Main Concepts"; that they have identified a legal basis for the processing of the special-category data; and that they inform the data subject about the processing[17] (see "Purpose limitation" in Part II, section "Principles").

Performing a '**data protection impact assessment' (DPIA) (**see "DPIA" within Part II section "Main tools and actions") is **compulsory if there is real risk of unauthorized profiling or automated decision-making.** Article 35(3)(a) of the GDPR states the need for the controller to carry out a DPIA in the case of a systematic and extensive evaluation of personal aspects relating to natural persons. This should be done for tools based on automated processing, including profiling, and for those on which decisions are based that produce legal effects concerning the natural person, or significantly affecting the natural person.

According to Article 37(1)(b)5 of the GDRP, an additional accountability requirement is the **designation of a data protection office (DPO)**, where the profiling or the automated decision-making is a core activity of the controller and requires regular and systematic monitoring of data subjects on a large scale. Controllers are also required to keep a **record of all decisions made by an AI system** as part of their accountability and documentation obligations (see Accountability section in Principles chapter). This should include whether an individual requested human intervention, expressed any views, contested the decision, and whether a decision has been altered as a result.[18]

Some additional actions that might be extremely useful to avoid automated decision-making include the following.[19]

- Consider the system requirements necessary to support a meaningful human review from the design phase.
- In particular, consider the interpretability requirements and effective user-interface design to support human reviews and interventions.
- Design and deliver appropriate training and support for human reviewers.
- Give staff the appropriate authority, incentives and support to address or escalate individuals' concerns and, if necessary, override the AI system's decision.

In any case, controllers should be aware that Member States are introducing some **concrete references to this issue in their national regulations**, using different tools to ensure adequate compliance.[20]

---

[17] Article 29 Working Party (2017) Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679, WP 251, p.15. European Commission, Brussels.

[18] ICO (2020) AI auditing framework: draft guidance for consultation, p.94-95. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (accessed 15 May 2020).

[19] Ibid. p.95.

Checklist: profiling and automated decision-making[21]

 The controllers have a legal basis to carry out profiling and/or automated decision-making, and document this in their data protection policy.

 The controllers send individuals a link to their privacy statement when they have obtained their personal data indirectly.

 The controllers explain how people can access details of the information that they used to create their profile.

 The controllers tell people who provide them with their personal data and how they can object to profiling.

 The controllers have procedures for customers to access the personal data input into their profiles, so they can review and edit for any accuracy issues.

 The controllers have additional checks in place for their profiling/automated decision-making systems to protect any vulnerable groups (including children).

 The controllers only collect the minimum amount of data needed and have a clear retention policy for the profiles that they create.

As a model of best practice

 The controllers carry out a DPIA to consider and address the risks when they start any new automated decision-making or profiling.

 The controllers tell their customers about the profiling and automated decision-making they carry out, what information they use to create the profiles, and where they get this information from.

 The controllers use anonymized data in their profiling activities.

 Those responsible guarantee the right to readability of algorithmic decisions.

 Decision-makers have a mechanism capable of notifying and explaining the reasons when a challenge to the algorithmic decision is not accepted due to lack of human intervention.

 The decision-makers have a model of human rights assessment in automated decision-making.

 Qualified human supervision is in place from the design phase onwards, in particular on

[20] Malgieri, G. (2018) Automated decision-making in the EU Member States laws: the right to explanation and other 'suitable safeguards' for algorithmic decisions in the EU national legislations. Available at: https://ssrn.com/abstract=3233611 or http://dx.doi.org/10.2139/ssrn.3233611 (accessed 2 May 2020).

[21] ICO (no date) Rights related to automated decision-making including profiling. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling/ (accessed 15 May 2020).

the interpretation requirements and the effective design of the interface, and the examiners are trained.

 Audits are conducted with respect to possible deviations from the results of inferences in adaptive or evolutionary systems.

 Certification of the AI system is being, or has been, carried out.

Additional information

Article 29 Working Party (2018) Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679. European Commission, Brussels. Available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053

ICO (2020) AI auditing framework: draft guidance for consultation, p.94-95. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

ICO (2019) Data Protection Impact Assessments and AI. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-data-protection-impact-assessments-and-ai/

Kosinski M., Stillwell, D. and Graepel, T. (2013) 'Digital records of behavior expose personal traits', *Proceedings of the National Academy of Sciences* 110(15): 5802-5805, DOI:10.1073/pnas.1218772110

Malgieri, G. (2018) Automated decision-making in the EU Member States laws: the right to explanation and other 'suitable safeguards' for algorithmic decisions in the EU national legislations. Available at: https://ssrn.com/abstract=3233611 or http://dx.doi.org/10.2139/ssrn.3233611

Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf

Selbst, A.D. and Powles, J. (2017) 'Meaningful information and the right to explanation', International Data Privacy Law 7(4): 233-242, https://doi.org/10.1093/idpl/ipx022

Wachter, S., Mittelstadt, B. and Floridi, L. (2017) 'Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation', International Data Privacy Law. Available at https://ssrn.com/abstract=2903469 or http://dx.doi.org/10.2139/ssrn.2903469

2 Technical robustness and safety

"A crucial component of achieving Trustworthy AI is technical robustness, which is closely linked to the principle of prevention of harm. Technical robustness requires that

AI systems be developed with a preventive approach to risks and in a manner such that they reliably behave as intended while minimizing unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured."

- *High-Level Expert Group on AI*[22]

2.1 Ethical principles and GDPR provisions

The High-Level Expert Group on AI splits the requirement for technical robustness and safety into four sub-components: (1) resilience to attack and security; (2) a fallback plan and general safety; (3) accuracy; and (4) reliability and reproducibility.

For easy referencing, this section mirrors this structure, while connecting these sub-components to legal (GDPR) requirements and recommendations. This is important, because while the GDPR requirements generally only apply when processing personal data, many practical AI systems are designed to produce a personalized result (i.e. recommender systems), and therefore have to process personal data at some point.

2.1.1 Resilience to attack and security

Resilience to attack should be a goal of all ICT systems, including AI systems. When processing personal data, Article 32 of the GDPR explicitly requires the implementation of appropriate technical and organizational measures to ensure data security (see 'Measures in support of confidentiality' in the 'Integrity and confidentiality' subsection of the "Principles" in Part II.

The required security measures depend on the likely impact of an AI system malfunction. These measures should also include steps taken to ensure the resilience of processing systems.[23] For certain types of AI system, the decision-making process may be particularly vulnerable to attack. For example, a malicious actor may create a misleading input to exploit the fundamental perception differences between humans and AI systems, as demonstrated by the example in Box 2.

Box 2. Example of the need for security in AI systems

An autonomous vehicle should automatically recognize street signs, by using on-board cameras and adjusting its speed accordingly. While AI algorithms based on Deep-Neural-Networks may excel in this task, special care must be taken to protect the system against targeted attacks by a malicious adversary. For example, small, targeted modifications to street signs could lead the AI system to mistake a stop sign for a speed limit sign, resulting

[22] High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI, p.16 and ff. European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 28 may 2020).

[23] Article 32(1)(b) of the GDPR.

in potentially dangerous situations. Meanwhile, the modification may appear as simple graffiti to the casual human observer. It is therefore of utmost importance to protect an AI system used for this purpose against such attacks, thereby increasing its resilience.[24]

Trained AI models can also be a valuable data source. Under certain circumstances, it may be possible to gain insights into the original input data using only the trained model.[25] Such 'information leakage' could be exploited by both internal and external actors. It is therefore important for controllers **to take measures to limit access to the model and underlying training data, and for all actor categories** (see "Measures in support of confidentiality" in the "Integrity and confidentiality" subsection of the "Principles" in Part II of these Guidelines).

Once trained, the resulting AI system may be used for very different purposes than originally intended. For example, face recognition AI system may be used to recognize and group photos containing a specific person within an online photo album. The same AI system could also be used to search the internet for photos of a specific person, potentially revealing sensitive personal details (i.e. using the photo location or capture context). This kind of multi-purpose use is often possible with AI systems, and it is up to the system designer to *predict possible unlawful processing of personal data and implement security measures that would prevent or minimize it.* This could be through measures such as restricting the usable data sources, or prohibiting certain usage patterns though licensing terms. Data protection legal framework may complement such restrictions, but is by no means a replacement for them.

2.1.2 Fallback plan and general safety

"To error is human but to really foul things up requires a computer."

- *Paul Ehrlich / Bill Vaughan*[26]

Like all ICT systems, AI systems may fail and provide incorrect results or predictions. However, in the case of AI systems, it may be particularly hard to explain why a particular (false) conclusion was reached in a tangible, human way. An example of undesirable behavior would be an AI system that makes decisions that significantly affect an individual (e.g. automatically denying a credit application). The GDPR requires controllers to implement suitable fallback plans protecting data subjects from such situations, including the right to contest an AI decision and to obtain a human intervention that considers the data subjects' point of view.[27] Such safeguards should be

[24] Eykholt, K. et al. (2018) 'Robust physical-world attacks on deep learning models', 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition 10.4.2018, arXiv:1707.08945.

[25] Fredrikson, M. et al. (2015) 'Model inversion attacks that exploit confidence information and basic countermeasures', CCS '15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, October 2015. Cornell University, Ithaca. Available at: https://rist.tech.cornell.edu/papers/mi-ccs.pdf (accessed 20 May 2020).

[26] The authorship the quote appears to be disputed, see i.e. https://quoteinvestigator.com/2010/12/07/foul-computer/#note-1699-18 (accessed 2 June 2020).

[27] Article 33(3) of the GDPR.

considered during the systems design. Even in cases where the GDPR does not explicitly require such a fallback plan, it is desirable for controllers to consider implementing one.

Controllers should also be aware of safety issues. New technologies often lead to new risks. It is important to be aware that protection of personal data depends on IT security measures and therefore risks related with personal data are those related with IT. Consequently, appropriate technical and organizational measures implemented in IT will provide data protection as is stipulated by GDPR, and those should be regularly tested and upgraded to prevent or minimize security risks. (see the subsection 'Main difference from other risks in the GDPR and from risks in IT security' in the 'Integrity and confidentiality' section in the 'Principles' chapter).

To assess these risks and derive appropriate safeguards, the GDPR requires a DPIA to be performed prior to processing when there is a high risk to the rights and freedoms of a natural person[28] **(**see "DPIA" within Part II section "Main tools and actions" of these Guidelines). The use of new technologies such as AI increases the likelihood that the processing falls into the high-risk category. Some national data protection agencies have issued directives requiring a DPIA when using certain AI algorithms.[29] In case of doubt, it is recommended that controllers perform a DPIA.[30]

2.1.3 Accuracy

High system accuracy is usually one of the design goals of AI systems. Many AI systems require accurate and reliable training data to achieve the best results. When processing personal data, keeping it up to date and correcting wrong inputs is also a legal requirement.[31] The data subject can also demand the rectification of inaccurate personal data.[32] AI systems should therefore be designed with the need for retraining in mind, during which data may not only be added - but also removed (see "Right to rectification" within Part II section "Data subject's rights" of these Guidelines and "Fairness, diversity and non-discrimination" within this Part III on AI, as well as, "Lawfulness, fairness and transparency principle" within Part II section "Principles").

In addition, the output of an AI system should not only be a result, but also a measure of how confident the system is that the result is correct. Such a measure is not only a technical indicator of the system's accuracy, but also a valuable indication of whether human intervention may be required (see the "Accuracy principle" section in the "Principles" within Part II of these Guidelines).

---

[28] Article 35(1) of the GDPR

[29] See, for example, the legal situation in Austria § 2(2)(4) DSFA-V.

[30] Article 29 Working Party (2017) WP248, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, p.8. European Commission, Brussels.

[31] Article 5(1)(d) of the GDPR.

[32] Article 16 of the GDPR.

2.1.4 Reliability and reproducibility

Many AI systems are designed with a specific use-case in mind. However, as stated, it may evolve over time and slowly drift away from the designers' original intentions. It is therefore important to document clearly the initial assumptions and conditions under which the AI system was intended to be used. For example, does the AI system expect a specific environment, or does the training set contain known biases? If an AI system is publicly available, the documentation of the system's reliability should be as well.

In addition to reliability, the reproducibility of an AI system's results is important. Not only is reproducibility a desirable technical property of an AI system (e.g. to investigate the reason for faulty results), it is also an important prerequisite for trust. If a result cannot be reproduced, its explainability - and therefore trust in the AI system - may suffer.

Checklist: technical robustness and safety[33]

Resilience to attack and security

- The controller assessed potential forms of attacks to which the AI system could be vulnerable.
- The controller considered different types and natures of vulnerabilities, such as data pollution, physical infrastructure and cyber-attacks.
- The controller put measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks.
- The controller verified how the system behaves in unexpected situations and environments.
- The controller considered to what degree the system could be dual-use. If so, the controller took suitable preventative measures against this (e.g. not publishing the research or deploying the system).

Fallback plan and general safety

- The controller ensured that the system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations (e.g. technical switching procedures or asking for a human operator before proceeding).
- The controller considered the level of risk raised by the AI system in this specific use case.
- The controller put any process in place to measure and assess risks and safety.

[33] This checklist has been adapted from the one elaborated by the High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI. European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 20 May 2020).

 The controller provided the necessary information in case of a risk to human physical integrity.

 The controller considered an insurance policy to deal with potential damage from the AI system.

 The controller identified potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse. Is there a plan to mitigate or manage these risks?

 The controller assessed whether there is a probable chance that the AI system may cause damage or harm to users or third parties. The controller assessed the likelihood, potential damage, impacted audience and severity.

 The controller considered the liability and consumer protection rules, and take them into account.

 The controller considered the potential impact or safety risk to the environment or to animals.

 The controller's risk analysis included whether security or network problems (e.g. cybersecurity hazards) could pose safety risks or damage due to unintentional behavior of the AI system.

 The controller estimated the likely impact of a failure of the AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (e.g. discrimination).

 The controller defined thresholds and put governance procedures in place to trigger alternative/fallback plans.

 The controller defined and test fallback plans.

Accuracy

 The controller assessed what level and definition of accuracy would be required in the context of the AI system and use case.

 The controller assessed how accuracy is measured and assured.

 The controller put in place measures to ensure that the data used is comprehensive and up to date.

 The controller put in place measures to assess whether there is a need for additional data, for example to improve accuracy or eliminate bias.

 The controller verified what harm would be caused if the AI system makes inaccurate predictions.

 The controller put in place ways to measure whether the system is making an unacceptable amount of inaccurate predictions.

 The controller put in place a series of steps to increase the system's accuracy.

Reliability and reproducibility

- The controller put in place a strategy to monitor and test if the AI system is meeting its goals, purposes and intended applications.
- The controller tested whether specific contexts or particular conditions need to be taken into account to ensure reproducibility.
- The controller put in place verification methods to measure and ensure different aspects of the system's reliability and reproducibility.
- The controller put in place processes to describe when an AI system fails in certain settings.
- The controller clearly documented and operationalize these processes for the testing and verification of the reliability of AI systems.
- The controller established mechanisms of communication to assure (end-)users of the system's reliability.

Additional information

Article 29 Working Party (2017) WP248, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679. European Commission, Brussels. Available at: https://ec.europa.eu/newsroom/document.cfm?doc_id=47711

Eykholt, K. et al. (2018) 'Robust physical-world attacks on deep learning models', 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition 10.4.2018, arXiv:1707.08945

Fredrikson, M. et al. (2015) 'Model inversion attacks that exploit confidence information and basic countermeasures', CCS '15: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, October 2015. Cornell University, Ithaca. Available at: https://rist.tech.cornell.edu/papers/mi-ccs.pdf

3 Privacy and data governance

"Closely linked to the principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy."

- High-Level Expert Group on AI[34]

[34] High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI, p.17. European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 28 May 2020).

3.1 Ethical principles

This requirement embeds three main different principles, namely:[35]

- **Privacy and data protection.** AI systems must guarantee privacy and data protection throughout a system's entire lifecycle. This includes the information initially provided by the user, as well as the information generated about the user over the course of their interaction with the system (e.g. outputs that the AI system generated for specific users or how users responded to particular recommendations). To allow individuals to trust the data-gathering process, it must be ensured that data collected about them will not be used to discriminate against them unlawfully or unfairly.
- **Quality and integrity of data.** The quality of the datasets used is paramount to the performance of AI systems. When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. This needs to be addressed prior to training with any given dataset. In addition, the integrity of the data must be ensured. Feeding malicious data into an AI system may change its behavior, particularly with self-learning systems. Processes and datasets used must be tested and documented at each step, such as planning, training, testing and deployment. This should also apply to AI systems that were not developed in-house but acquired elsewhere.
- **Access to data.** In any given organization that handles individuals' personal data (internal documents/policies stipulating who and under what conditions may access personal data including organizational and technical measures of access control, must be in place. Only duly qualified personnel with the competence and need to access individual's personal data should be allowed to do so. Also, all personnel that are granted access has to sign confidentiality statement.

3.2 GDPR provisions

GDPR refers to the processing of data subjects' personal data. There are some provisions that are particularly relevant to privacy and data governance. Since the quality and integrity of data and access to data are analyzed in the previous section, our focus here is on four concepts that are extremely relevant to guaranteeing adequate data governance. These are: (1) purpose limitation; (2) lawfulness; (3) data minimization; and (4) fairness, a broad principle that requires protecting data subjects' rights.

It is pointless to talk about data protection if the processing is not lawful, and a specified and explicit purpose is a prerequisite for lawful processing. However, even if the processing is permissible (i.e. lawful and legitimate), data protection remains impossible to implement if the purposes of processing are unclear. Moreover, processing is not lawful if it is not related to the purposes for which the data were collected. Therefore, the purpose limitation principle is directly connected with data governance.

[35] Ibid., p.15 and ff.

Meanwhile, data minimization is key to protecting privacy. The best way to ensure that "data collected about [the data subjects] will not be used to unlawfully or unfairly discriminate against them"[36] is to minimize the amount and range of personal data collected. Lastly, adequate implementation of data subjects' rights, as embedded in the GDPR, is essential to empower them and strengthen the data governance framework.

3.2.1 Purpose limitation

The purpose limitation principle limits the use of personal data to the original purpose(s), or those purposes that are compatible with it. However, AI development requires data to be reused quite often. Moreover, it might happen that the AI tool re-uses the data automatically (this certainly happens in the case of deep learning). These situations create tension between the AI training techniques and the principle of purpose limitation (see "Purpose limitation principle" within Part II section "Principles" of these Guidelines).

In order to avoid unlawful data processing, controllers using AI systems **should determine the purpose of the processing "at the outset of its training or deployment, and perform a re-assessment of this determination should the system's processing throw up unexpected results**, since it requires that personal data only be collected for "specified, explicit and legitimate purposes" and not be used in a way that is incompatible with the original purpose"[37] (see the "Data protection by design and by default" section in "Main Concepts", within Part II of these Guidelines).

The **re-use of data** in the development of an AI tool entails deeply challenging issues in terms of purpose limitation. AI systems process personal data in various stages and for a variety of purposes. As a result, a controller may fail to distinguish each distinct processing operation and process data for purposes others than those for which they were initially collected. Controllers should be particularly concerned about these situations since they could lead to a failure to comply with the data protection principle of lawfulness[38] (see the "Use for incompatible purposes" subsection in "Purpose limitation principle'" section of the "Principles" within Part II of these Guidelines).

Controllers must consider that the identification of the appropriate lawful basis **is tied to principles of fairness and purpose limitation** (see "Lawfulness, fairness and

---

[36] High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI, p.17. European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 28 May 2020).

[36] Ibid., p.15 and ff.

[37] CIPL (2020) Artificial intelligence and data protection: how the GDPR regulates AI. Centre for Information Policy Leadership, Washington DC / Brussels / London, p.6. Highlighted by the author. Available at: www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf (accessed 17 May 2020).

[38] ICO (2020) Guidance on the AI auditing framework: draft guidance for consultation. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (accessed 15 May 2020).

transparency principle" within Part II section "Principles" of these Guidelines).[39] They must select the legal basis that most closely reflects the true nature of their relationship with the individual and the purpose of the processing. This decision is key, since changing the legal grounds for processing is impossible if there are not substantial reasons that justify it, due to the purpose limitation principle. If the AI developers are planning to use a dataset at different stages (e.g. training, validation or deployment), they should consider these steps as having distinct and separate purposes.[40] Moreover, **they must consider the type of relationship they hold with the data subject.** For instance, consent might be an appropriate lawful basis for processing if there is on-going contact with the data subjects and controllers are able to obtain successive consents for different uses or are able to obtain consent for several processings from data subject before processing starts. However, in the case of AI, it is often hard to keep this type of relationship, since AI is often built by aggregating and merging big datasets.

Last but not least, controllers should be aware that for processing of personal data for scientific, historical research or statistical purposes, Union or Member State law or rules may provide derogations from data subjects' rights stipulated in Art. 15,16,18,21- Therefore processing of those data for purposes other than those for which they were initially collected should be lawful as long as appropriate technical and organizational measures are in place, in particular data minimization. (see the "Data protection and scientific research" within "Main Concepts" in Part Ii of these Guidelines).

Checklist: purpose limitation[41]

- The controllers have clearly identified their purpose or purposes for processing.
- The controllers have documented those purposes.
- The controllers include details of their purposes in the privacy information for individuals.
- The controllers regularly review their processing and, where necessary, update their documentation and privacy information for individuals.
- If the controllers plan to use personal data for a new purpose other than a legal obligation or function set out in law, they check that this is compatible with their original

[39] EDPB (2018) Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects, Adopted on 9 April 2019, p.6. European Data Protection Board, Brussels. Available at: https://edpb.europa.eu/sites/edpb/files/consultation/edpb_draft_guidelines-art_6-1-b-final_public_consultation_version_en.pdf (accessed 15 May 2020).

[40] ICO (2020) Guidance on the AI auditing framework: draft for consultation. 2020. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (accessed 15 May 2020).

[41] ICO (no date) Principle (b): purpose limitation. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/purpose-limitation/ (accessed 17 May 2020).

purpose or they get specific consent for the new purpose.

Additional information

Article 29 Data Protection Working Party (2013) Opinion 03/2013 on purpose limitation. European Commission, Brussels. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

CIPL (2020) Artificial intelligence and data protection: how the GDPR regulates AI. Centre for Information Policy Leadership, Washington DC / Brussels / London. Available at: www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf

EDPB (2018) Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects, Adopted on 9 April 2019, p.6. European Data Protection Board, Brussels. Available at: https://edpb.europa.eu/sites/edpb/files/consultation/edpb_draft_guidelines-art_6-1-b-final_public_consultation_version_en.pdf

ICO (2020) Guidance on the AI auditing framework: draft guidance for consultation. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

ICO (no date) Principle (b): purpose limitation. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/purpose-limitation/

3.2.2 Lawfulness

Lawfulness is an essential principle in terms of data protection. It implies that controllers shall ensure that they have a **legal basis for processing personal data**. **If this is not the case,** the processing **must not be carried out**.[42] In general, and including special categories data, legal bases for processing are described in Article 6 and Article 9 of the GDPR. In the case of AI, the legal bases that are usually invoked to justifying processing are: consent; legitimate interest; contractual necessity; and legal obligation or vital interest. Processing for public interest can also be a legal ground, but we will not focus on it here since we address this topic widely in the section "Data protection and scientific research" in the "Main Concepts" section of Part II of these Guidelines. Therefore, we will focus on the four legal grounds listed.

[42] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial: Una introducción, p.20. Agencia Espanola Proteccion Datos, Madrid. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 15 May 2020).

a) Consent

Data processing is often based on consent provided by the data subjects. However, consent does not fit well with the essential nature of most AI developments, due to one simple fact: consent is, by nature, linked to a well-defined, concrete purpose.[43] In the case of AI, the use of big data, and the aggregation, sharing or repurposing actions that are often performed creates a scenario that does not fit the underlying principles of the concept of consent and the purpose limitation principle (see "Purpose limitation principle" within Part II section "Principles" of these Guidelines).

Consent can be a useful legal basis for data processing for AI development, especially if controllers have a **direct relationship with the subject who provides the data to be used for training, validating and deploying the model**.[44] For instance, if the AI tool aims to provide diagnoses of pneumonia, and physicians obtain data from patients at their healthcare institution, consent might serve well as a legal basis for processing. However, if processing involves the use of a complex AI tool that may have further uses of the data (e.g. profiling and automated decision-making might happen inadvertently, data are likely to be inferred during processing, such inferred data can be used for various purposes, etc.), it is difficult to see how a single consent could justify all such processing. To this end, controllers must take good care of the guidelines on consent provided by the Article 29 Working Party[45] .

Within the framework of scientific research (see the "Data protection and scientific research" section in the "Main Concepts" in Part II of these Guidelines), the GDPR provides specific derogation from consent attributes, allowing controllers to make use of **broad consent** as a legal basis for processing. Broad consent must be understood in connection with Recital 33 of the GDPR, which states that it "is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognized ethical standards for scientific research."

However, broad consent is not a kind of blanket or equivocal to open consent. It is an **exceptional tool that can only be acceptable if several conditions apply.** If broad consent is used for special categories of data, controllers should ensure that their national regulation allows for this. They should also be aware of the safeguards that should be implemented. Proportionality between the aim of research and the use of special categories of data must be guaranteed. Furthermore, controllers must ensure that their Member States regulation do not protect genetic, biometric and health data by

---

[43] International Bioethics Committee (2017) Report of the IBC on big data and health, p.20. UNESCO. Available at: http://unesdoc.unesco.org/images/0024/002487/248724e.pdf (accessed 13 March 2020).

[44] ICO (no date) How do we apply legitimate interests in practice? Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/legitimate-interests/how-do-we-apply-legitimate-interests-in-practice/ (accessed 15 May 2020). Furthermore, the assessment of the nature of this relationship must include an investigation of the balance of power between the data subject and the data controller.

[45] Article 29 Working Party (2018) Guidelines on consent under Regulation 2016/679. European Commission, Brussels, p.29. Available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051 (accessed 5 May 2020).

introducing further conditions or limitations, since they are allowed to do so by the GDPR.

Furthermore, whenever broad consent is used to achieve the research purpose, there are some essential measures that should be considered to compensate for the abstract definition of the research purposes. Adherence to the recognized ethical standards of scientific research, according to Recital 33 of the GDPR, seems particularly relevant to this purpose.

Box 3: Broad consent and additional safeguards

The German DPA recently listed some additional safeguards to be implemented in the case of broad consent.[46] These are:

1. Safeguards to ensure transparency:

- Utilization of usage regulations or research plans that illustrate the planned working methods and questions that are to be the subject of the research project.
- Assessment and documentation of the question why in this particular research project a more detailed specification of the research purposes is not possible.
- Set up web presences to inform study participants about ongoing and future studies.

2. Safeguards to build trust:

- Positive vote of an ethics committee before use of data for further research purposes.
- Assessment of whether it is possible to work with a dynamic consent or whether a data subject can object before the data might be used for new research questions.

3. Security safeguards:

- No data transfers to third countries with a lower level of data protection
- Additional measures regarding data minimization, encryption, anonymization, or pseudonymization
- Implementation of specific policies to limit access to personal data.

In any case, research participants must be given the **possibility to withdraw their consent,** and opt in or out of certain research and parts of research, and be assured that

[46] DSK, Beschluss der 97. Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder zu Auslegung des Begriffs „bestimmte Bereiche wissenschaftlicher Forschung“ im Erwägungsgrund 33 der DS-GVO 3. April 2019, at: www.datenschutzkonferenz-online.de/media/dskb/20190405_auslegung_bestimmte_bereiche_wiss_forschung.pdf (accessed 20 May 2020). The English translation comes from a nice summary of the measures that can be consulted here: www.technologylawdispatch.com/2019/04/privacy-data-protection/german-dpas-publish-resolution-on-concept-of-broad-consent-and-the-interpretation-of-certain-areas-of-scientific-research/

their rights are safeguarded by adherence to the ethical standards of scientific research.[47] Sometimes, this might cause harm the AI solution or oblige the controllers to perform complex actions. Therefore, controllers should consider if alternative legal grounds could serve them better to develop the tool while respecting the law.

In summary, controllers should be **cautious when using consent as a legal ground to justify data processing**, since consent does not invalidate their responsibilities regarding the fairness, necessity and proportionality of the processing.[48] Furthermore, in the case of AI using Big Data, it is often hard to justify that consent fulfils all necessary requirements: freely given, specific, informed and unambiguous, and a clear affirmative act on the part of the data subject. In general, the more things that AI developers want to do with the data, the more difficult it is to ensure that consent is genuinely specific and informed. This should all be considered when selecting consent as a legal ground for data processing.

Box 4. Consent as a legal basis: the OkCupid case

In 2016, a group of Danish researchers published a dataset of about 70,000 users. These data had been obtained from the online dating site OkCupid[49] and included data categories such as usernames, age, gender, location, what kind of relationship (or sex) the data subjects were interested in, their personality traits, etc.

The researchers considered that the mere fact that these data were publicly available (on the users' dating profiles) served as legal grounds for further processing. This is an excellent example of the terrible consequences of the argument that "the data is already public". Data subjects had their personal data processed, and very sensitive information exposed to the public, without their consent.

Unfortunately, this association between public and open data is still too extensive. Researchers should be aware that consent provided for one concrete processing does not serve as legal grounds for further processing, and that 'publicly available' is not equivocal to 'open data'; that is, data and content that can be freely used, modified and share by anyone for any purpose, as defined by the Open Data Institute.[50]

[47] Kuyumdzhieva, A. (2018) 'Ethical challenges in the digital era: focus on medical research', pp.45-62 in: Koporc, Z. (ed.) *Ethics and integrity in health and life sciences research*. Emerald, Bingley.

[48] Article 29 Working Party (2018) Guidelines on consent under Regulation 2016/679. WP259. European Commission, Brussels, p.3. Available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051 (accessed 15 May 2020).

[49] www.okcupid.com (accessed 5 May 2020).

[50] http://opendefinition.org/ (accessed 5 May 2020).

Checklist: consent

- The controllers have checked that consent is the most appropriate legal basis for processing.
- The controllers request the consent of the interested parties in a free, specific, informed and unequivocal manner.
- Broad consent is used only when it is difficult or improbable to foresee how this data will be processed in the future.
- Broad consent used for processing of special categories of data is compatible with national regulations.
- Where broad consent is used, the data subjects are given the opportunity to withdraw their consent and to choose whether or not to participate in certain research and parts of it.
- Controllers have a direct relationship with the subject who provides the data to be used for training, validation and deployment of the IA model.
- There is no power imbalance between controllers and data subjects.
- The controllers ask people to positively opt in.
- The controllers do not use pre-ticked boxes or any other type of default consent.
- The controllers use clear, plain language that is easy to understand.
- The controllers specify why they want the data and what they are going to do with it.
- The controllers give separate distinct ('granular') options to consent separately to different purposes and types of processing.
- The controllers tell individuals they can withdraw their consent and how to do so.
- The controllers ensure that individuals can refuse to consent without detriment.
- The controllers avoid making consent a precondition of a service.

b) Legitimate interest

The use of legitimate interest as a legal ground for processing for AI development is applicable, provided that the result of the balancing test justifies it (see "Legitimate interest and balancing test" within Part II section "Main actions and tools" of these Guidelines). This may imply defining the objective of the AI's processing at the outset, and ensuring that the original purpose of the processing is re-evaluated if the AI system provides an unexpected result, so that either the legitimate interests pursued can be identified, or that valid consent can be collected from individuals.[51] The balancing test **should be adequately documented in the records of processing.** However, in some cases, legitimate interest might not serve well for AI processing purposes. For example,

[51] CIPL (2020) Artificial intelligence and data protection. How the GDPR regulates AI. Centre for Information Policy Leadership, Washington, DC/Brussels/London, p.5. Available at: www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf (accessed 15 May 2020).

if controllers plan to gather a considerable amount of personal data 'just in case', they should not consider legitimate interest as a legal ground for the data processing, since the balancing between the need for processing and the possible impacts of the processing on people would hardly justify it.[52]

Checklist: legitimate interest as a legal basis

- The controllers have checked that legitimate interest is the most appropriate basis.
- The controllers understand their responsibility to protect individuals' interests.
- The controllers keep a record of the decisions made and the reasoning behind them, to ensure that they can justify their decision.
- The controllers have identified the relevant legitimate interests.
- The controllers have checked that the processing is necessary and there is no less intrusive way to achieve the same result.
- The controllers have done a balancing test and are confident that the individual's interests do not override those legitimate interests.
- The controllers only use individuals' data in ways they would reasonably expect, unless the controllers have a very good reason.
- The controllers are not using people's data in ways they would find intrusive, or which could cause them harm, unless the controllers have a very good reason.
- If the controllers process children's data, they take extra care to make sure they protect the children's interests.
- The controllers have considered safeguards to reduce the impact, where possible.
- The controllers have considered whether they can offer an opt out.
- The controllers have considered whether they also need to conduct a DPIA.

c) Performance of a contract

Performance of a contract to which the data subject is party, or in order to take steps at the request of the data subject prior to entering into a contract, might serve as a legal ground for processing, if using AI is objectively necessary to any of these purposes. This could be the case for developers who hire subjects to make use of their personal data in the training stage of the system. It could also be the case that the controller, who provides a service to interested third parties that includes the IA solution, uses the data of these subjects in the framework of the service contract.[53] However, this legal ground should not be used for different purposes (such as system improvement or similar) according to the principle of purpose limitation (see "Purpose limitation principle"

[52] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.22. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 15 May 2020).

[53] Ibid., p.20.

within Part II section "Principles" of these Guidelines), since data used to perform the contract are not necessary for those alternative aims.[54] Thus, controllers can process the data that are intrinsically needed for the performance of a contract under the umbrella of this legal basis if they are objectively necessary to perform the contract, but not for other purposes.[55] To sum up, it seems hard to see how the performance of a contract might serve as legal basis for AI research and innovation.

d) Legal obligation or vital interest

According to Article 6(1)(d) of the GDPR, data can be processed if it is "necessary in order to protect the vital interests of the data subject or of another natural person". Equally, processing is lawful if it is "necessary for compliance with a legal obligation to which the controller is subject" (Article 6(1)(c). If we talk about special categories of data, then there are alternative legal grounds for processing, as expressed in Article 9.2. It is again **difficult to imagine a single case where any of these bases could provide a legal ground for training an AI system** at this moment, even though revisions of existing regulations at the national and European level may change this in the future. In any case, for the training of potentially life-saving AI systems, it would be better to rely on other legal bases, such as consent or public interest.[56]

Box 5. Examples of vital interest as a legal ground for data processing by an AI tool

Imagine that, during the COVID-19 pandemic, an organization develops an AI tool able to diagnose the disease using radiology. In such cases, data pertaining to patients could be processed on the basis of vital interest, as stated by Article 9(2)(c) of the GDPR. However, alternative legal grounds, such as substantial public interest (Article 9(2)(g) or (i) might be more appropriate.

Additional information

AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.20 Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf

Article 29 Working Party (2014) Opinion 6/2014 on the notion of legitimate interests of the

[54] Article 29 Data Protection Working Party (2014) Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. European Commission, Brussels, pp.16-17. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf (accessed 16 May 2020).

[55] EDPB (2019) Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects. European Data Protection Board, Brussels, p.14. Available at: https://edpb.europa.eu/sites/edpb/files/consultation/edpb_draft_guidelines-art_6-1-b-final_public_consultation_version_en.pdf (accessed 15 May 2020).

[56] Article 29 Working Party (2014) Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. European Commission, Brussels, p.20. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf (accessed 15 May 2020).

controller under Article 7 of Directive 95/46. European Commission, Brussels. Available at: www.dataprotection.ro/servlet/ViewDocument?id=1086

CIPL (2020) Artificial intelligence and data protection. How the GDPR regulates AI. Centre for Information Policy Leadership, Washington, DC / Brussels / London. Available at: www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf

EDPB (2019) Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects. European Data Protection Board, Brussels. Available at: https://edpb.europa.eu/sites/edpb/files/consultation/edpb_draft_guidelines-art_6-1-b-final_public_consultation_version_en.pdf

EDPB (2020) Guidelines 05/2020 on consent under Regulation 2016/679 Version 1.1 Adopted on 4 May 2020. Available at: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_guidelines_202005_consent_en.pdf

EDPS (2017) Necessity toolkit. European Data Protection Supervisor, Brussels. Available at: https://edps.europa.eu/data-protection/our-work/publications/papers/necessity-toolkit_en

Further reading about legitimate interest, with practical cases and several references to the rulings by the Court of Justice of the European Union can be found in the following documents.

Future of Privacy Forum (no date) Processing personal data on the basis of legitimate interests under the GDPR. European Judicial Training Network, Brussels. Available at: www.ejtn.eu/PageFiles/17861/Deciphering_Legitimate_Interests_Under_the_GDPR%20(1).pdf

ICO (no date) How do we apply legitimate interests in practice? Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/legitimate-interests/how-do-we-apply-legitimate-interests-in-practice/

ICO (no date) Lawful basis for processing. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/

Kuyumdzhieva, A. (2018) 'Ethical challenges in the digital era: focus on medical research', pp. 45-62 in: Koporc, Z. (ed.) Ethics and integrity in health and life sciences research. Emerald, Bingley.

Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf

3.2.3 Data minimization

The data minimization principle stipulates that personal data should be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed".[57] In the AI context, this means, in the first instance, that **controllers should**

[57] Article 5(1)(c) of the GDPR.

avoid using personal data if it is not necessary; that is, if the objective that the controller is aimed at can be obtained without processing personal data (see the "Lawfulness, fairness and transparency" within "Principles" in Part II of these Guidelines). Indeed, sometimes personal data can be substituted with non-personal data without affecting the research purposes. In such circumstances, the use of anonymized data is compulsory, according to Article 89.1 of the GDPR.

If anonymization is not possible, controllers should at least try to work with pseudonymized data. Ultimately, each controller needs to define which personal data are actually needed (and which are not) for the purpose of the processing, including the relevant data retention periods. Indeed, controllers must keep in mind that the necessity of processing must be proven in the case of most legal bases - including all those bases stated in Article 6 of the GDPR except consent, and most of the bases included in Article 9(2) regarding special categories of data. In other words, for the majority of legal bases for processing personal data, both data minimization and lawfulness principles require controllers to ensure that AI development cannot be done without using personal data.

The concept of necessity is, however, complex, and has an independent meaning in European Union law.[58] In general, it requires that processing is a targeted and proportionate way of achieving a specific purpose. It is not enough to argue that processing is necessary because controllers have chosen to operate their business in a particular way. They must be able to demonstrate that the processing is **necessary for the objective being pursued** and is **less intrusive than other options** for achieving the same goal; not that it is a necessary part of their chosen methods.[59] If there are realistic, less intrusive alternatives, the processing of personal data is not deemed necessary.[60]

Therefore, the data minimization principle requires AI developers to opt for those tools whose development involves minimal use of personal data compared to the available alternatives. Once this has been reached, specific processes should be in place to exclude unnecessary personal data being collected and/or transferred, reduce data fields and provide for automated deletion mechanisms.[61] Data minimization may be especially complex in the case of deep learning, where discrimination by features might be impossible. Therefore, if alternative solutions might bring the same outcomes, deep learning should better be avoided.

---

[58] See CJEU, Case C-524/06, Heinz Huber v Bundesrepublik Deutschland, 18 December 2008, para. 52.

[59] EDPS (2017) Necessity toolkit: assessing the necessity of measures that limit the fundamental right to the protection of personal data, p.5. European Data Protection Supervisor, Brussels. Available at: https://edps.europa.eu/data-protection/our-work/publications/papers/necessity-toolkit_en (accessed 15 May 2020); ICO (no date) Lawful basis for processing. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/ (accessed 15 May 2020).

[60] See CJEU, Joined Cases C-92/09 and C-93/09, Volker und Markus Schecke GbR and Hartmut Eifert v Land Hessen, 9. November 2010.

[61] ENISA (2015) Privacy by design in big data. European Union Agency for Cybersecurity, Athens / Heraklion, p.23. Available at: www.enisa.europa.eu/publications/big-data-protection (accessed 28 May 2020).

Further, the CIPL notes that "what personal data is considered 'necessary' varies depending on the AI system and the objective for which it is used, but the governance of the GDPR in this area should prevent the perfect from being the enemy of the good for AI designers – the fact that the personal data must be limited does not mean that the AI system itself becomes useless, especially since not all AI systems need to provide a precise output."[62] In order to determine precisely the range and amount of personal data needed, **having an expert able to select relevant features becomes extremely important.** This should significantly reduce the risk to data subjects' privacy – without losing quality.

There is an efficient tool to regulate the amount of data gathered and increase it only if it seems necessary: the **learning curve**.[63] The controller should start by gathering and using a restricted amount of training data, and then monitor the model's accuracy as it is fed with new data. This will also help a controller to avoid the 'curse of dimensionality'; that is, "a poor performance of algorithms and their high complexity associated with data frame having a big number of dimensions/features, which frequently make the target function quite complex and may lead to model overfitting as long as often the dataset lies on the lower dimensionality manifold."[64]

Some additional measures related to the minimization principle include:

- limit the extension of the data categories (e.g. names, physical and addresses, fields about their health, work situation, beliefs, ideology, etc.)
- limit the degree of detail or precision of the information, the granularity of the collection in time and frequency, and the age of the information used
- limit the extension in the number of interested parties of those who treat the data
- limit the accessibility of the different categories of data to the staff of the controller/manager or even the end-user (if there are data from third parties in the AI models) at all stages of the processing.[65]

Of course, adopting these measures might require a huge effort in terms of data unification, homogenization, etc., but it will contribute towards implementing the principle of data minimization in a much more efficient way.[66]

---

[62] CIPL (2020) Artificial intelligence and data protection: how the GDPR regulates AI. Centre for Information Policy Leadership, Washington DC / Brussels / London, p.13. Available at: www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf (accessed 15 May 2020).

[63] See: www.ritchieng.com/machinelearning-learning-curve/ (accessed 28 May 2020).

[64] Oliinyk, H. (2018) Why and how to get rid of the curse of dimensionality right (with breast cancer dataset visualization). Towards Data Science, 20 March. Available at: https://towardsdatascience.com/why-and-how-to-get-rid-of-the-curse-of-dimensionality-right-with-breast-cancer-dataset-7d528fb5f6c0 (accessed 15 May 2020).

[65] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.39-40. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 15 May 2020).

Finally, it is useful to remember that controllers should **avoid keeping long databases of historical data,** for example beyond the period required for normal business purposes, or to fulfil legal obligations, or simply because their analytic tool is able to produce a large number of data and its storage capacity makes this possible. Instead, companies using big data must enforce appropriate retention schedules (see the "Storage limitation" section in the "Principles", Part II of these Guidelines).

Box 6. An example of the benefits of data minimization in AI

An AI tool developed by the Norwegian tax administration to filter tax returns for errors tested five hundred variables in the training phase. However, only thirty were included in the final AI model, as they proved the most relevant to the task at hand. It is likely that the tool developers could have avoided collecting so many personal data, if they had performed a better selection of the relevant variables at the beginning of the development process.

Source: Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf

Checklist: data minimization

- The controllers have ensured that they only use personal data if needed.
- The controllers have considered the proportionality between the amount of data and the accuracy of the AI tool.
- The controllers periodically review the data they hold, and delete anything they do not need.
- The controllers at the training stage of the AI system debug all information not strictly necessary for such training.
- The controllers check if personal data are processed at the distribution stage of the AI system and delete them unless there is a justified need and legitimacy to keep them for other compatible purposes.

Additional information

ENISA (2015) Privacy by design in big data. European Union Agency for Cybersecurity, Athens / Heraklion, p.23. Available at: www.enisa.europa.eu/publications/big-data-protection

[66] Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf (accessed 15 May 2020).

ICO (no date) Principle (c): data minimization. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/data-minimisation/

Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf

Pure Storage (2015) Big data's big failure: the struggles businesses face in accessing the information they need. Pure Storage, Mountain View, CA. Available at: http://info.purestorage.com/rs/225-USM-292/images/Big%20Data%27s%20Big%20Failure_UK%281%29.pdf

3.2.4 Fairness with respect to data subjects' rights

Fairness is an essential concept in data protection (see the "Fairness" subsection in the "Lawfulness, fairness and transparency" section of the "Principles", as well as "Data subjects' rights", both within Part II of these Guidelines), one that can hardly be reached without an awareness that the development of AI tools can damage data subjects' interests, rights and freedoms. This is why it makes sense to ensure that adequate safeguards are implemented not only to avoid unfair consequences, but also to provide data subjects with enforceable rights that ensure adequate protection against unfair processing.

In this section, we explore how the key rights recognized by the GDPR apply to the AI development framework. To this purpose, we will concentrate on some rights that are particularly relevant in this area: (a) right to information (b) the right to access; (c) the right to data portability; (d) the right to rectification; and (e) the right to erasure; and (f) the right to object.

However, before considering this, researchers should check whether their research **could be considered as scientific research under Article 89 of the GDPR**. This is extremely important: if this is the case, EU or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 (address, rectification, restriction and object – and, indirectly, portability). These are subject to the conditions and safeguards referred to in paragraph 1 of Article 89, insofar as such rights are likely to render impossible, or seriously impair, the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes (see the "Data protection and scientific research" section in the "Main Concepts", Part II of these Guidelines).

a) Right to information

According to Article 13 of the GDPR, before processing personal data, the controller should provide the data subjects with complete information about the processing and their rights in an understandable format. If "the controller intends to further process the

personal data for a purpose other than that for which the personal data were collected, the controller shall provide the data subject prior to that further processing with information on that other purpose and with any relevant further information" (Article 13(3)).

However, the controllers are exempted of providing information to the data subjects if: the provision of such information proves impossible; would involve a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes; or the obligation to provide information is likely to render impossible or seriously impair the achievement of the objectives of that processing. Under such circumstances, controllers should take appropriate measures to protect the data subject's rights, freedoms and legitimate interests, including making the information publicly available. This derogation, however, is conditioned to the adoption of the safeguards imposed by Article 89 (see the "Data protection and scientific research" within Part II section "Main Concepts" of these Guidelines).

b) Right to access

Data subjects' right to access their data must be **guaranteed in all the steps of an AI tool's life cycle.** Controllers are encouraged to implement adequate technical measures to ensure that such access is easily reached by the data subject. Indeed, Article 15 of the GDPR gives the data subject the right to obtain details of any personal data used for profiling, including the categories of data used to construct a profile. Furthermore, pursuant to Article 15(3), the controller has a duty to make available the data used as input to create the profile, as well as access to information on the profile and details of which segments the data subject has been placed into. Similarly, Recital 63 of the GDPR states that "[w]here possible, the controller should be able to provide remote access to a secure system which would provide the data subjects with direct access to their personal data". **This includes observed, derived and inferred data**.[67]

Box 7. The inferred data issue

One of the most urgent issues we face in the realm of AI is the concrete status of inferred data. These are data that are not provided by the data subjects, but 'attributed' to them from available data, either from the same persons or from other persons. Sometimes, these inferred data provide information about an identifiable person. Regardless of whether this information is accurate or not, these data must be considered as personal data and the GDPR therefore applies to them. As a result, data subjects' rights should be strictly respected, including the right to access to such data.[68] However,

[67] ICO (2014) Big data and data protection. Information Commissioner's Office, Wilmslow, pp.99-10. Available at: https://rm.coe.int/big-data-and-data-protection-ico-information-commissioner-s-office/1680591220 (accessed 28 May 2020).

[68] See: Custers, B. (2018) 'Profiling as inferred data. Amplifier effects and positive feedback loops', pp.112-115 in Bayamlıoğlu, E. et al. (eds) *Being profiled: cogitas ergo sum. 10 years of profiling the European Citizen*. Amsterdam University Press, Amsterdam. DOI 10.5117/9789463722124/CH19.

as discussed elsewhere in these Guidelines, inferred data are not included in the right to portability (see "Right to portability" within Part II section "Data subject's rights" of these Guidelines).

One of the main problems embedded in AI and big data processing is that the right to access may, at times, collide with the interest of a company in keeping its commercial secrets. Indeed, Recital 63 of the GDPR provides some protection for controllers who are unwilling to unveil trade secrets or intellectual property, which may be particularly relevant in relation to profiling.[69] However, AI developers **cannot rely on the protection of their trade secrets as an excuse to deny access or refuse to provide information** to the data subjects**.** Instead, organizations need to find pragmatic solutions.[70]

The right to access might be more or less applicable, depending on the lifecycle stage at which the AI development is. For instance, providing access to training data to an individual data subject, might be hard since they usually only include information relevant to predictions (e.g. past transactions, demographics, location), but not contact details or unique customer identifiers. Moreover, they are often pre-processed to make them more amenable to machine learning algorithms. However, this does not mean at all that these data can be considered as anonymized. Thus, they continue to be personal data. For instance, in the case of a purchase prediction model, the training might include a pattern of purchases unique to one customer. In this example, if a customer were to provide a list of their recent purchases as part of their request, the organization may be able to identify the portion of the training data that relates to that individual.

Under such circumstances, **controllers must respond to data subjects' request to gain access to their personal data, assuming they have taken reasonable measures to verify the identity of the data subject, and no other exceptions apply.** And, as the ICO states, "requests for access, rectification or erasure of training data should not be regarded as manifestly unfounded or excessive just because they may be harder to fulfil or the motivation for requesting them may be unclear in comparison to other access requests an organization typically receives".[71] However, **organizations do not have to collect or maintain additional personal data to enable identification of data subjects in training data for the sole purposes of complying with the regulation.** If the controllers cannot identify a data subject in the training data, and the data subject

Available at: https://ssrn.com/abstract=3466857 or http://dx.doi.org/10.2139/ssrn.3466857 (accessed 28 May 2020).

[69] A29WP (2016) Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679. European Commission, Brussels, p.17. Available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053 (accessed 28 May 2020).

[70] Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo, p.19. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf (accessed 28 May 2020).

[71] ICO (2019) Enabling access, erasure, and rectification rights in AI systems. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/ (accessed 28 May 2020).

cannot provide additional information that would enable their identification, they are not obliged to fulfil a request that is not possible to satisfy.[72]

Checklist: right to access[73]

Preparing for subject access requests

- The controllers know how to recognize a subject access request and they understand when the right of access applies.
- The controllers understand that the right of access is to be applied at each stage of the life cycle of the AI solution, if it uses personal data.
- The controllers have a policy for how to record requests they receive verbally.
- The controllers understand when they can refuse a request and are aware of the information they need to provide to individuals when doing so.
- The controllers understand the nature of the supplementary information they need to provide in response to a subject access request.

Complying with subject access requests

- The controllers have processes in place to ensure that they respond to a subject access request without undue delay and within one month of receipt.
- The controllers are aware of the circumstances in which they can extend the time limit to respond to a request.
- The controllers understand that there is a particular emphasis on using clear and plain language if they are disclosing information to a child.
- The controllers understand what they need to consider if a request includes information about others.
- The controllers understand how to apply the right to access in training stages.

Additional information

Article 29 Working Party (2014) Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. European Commission, Brussels. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf

ICO (2013) Big data, artificial intelligence, machine learning and data protection.

[72] Ibid.

[73] ICO (no date) Right of access. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-of-access/ (accessed 28 May 2020).

Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

ICO (no date) Right of access. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-of-access/

Norwegian Data Protection Authority. (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf

c) Right to data portability

Article 20 of the GDPR created a new right: the right to data portability.[74] This right provides data subjects with control of the use of their data by redirecting it where it is most useful (see the "Right to data portability" section in the "Data Subject Rights" within Part II of these Guidelines). However, right to data portability might be hard to implement in the AI arena, for several reasons. One must keep in mind the cost and feasibility of providing extremely large, complex datasets accumulated over many years. This could make it hard for a company to fulfil their right to data portability requirements.

There are different types of personal data that a machine learning system can process. According to the Article 29 Data Protection Working Party, some categories of data are linked to the right of data portability, namely: personal data concerning the data subject and data which they have provided to a controller. In general, the term 'provided by the data subject' must be interpreted **broadly**. Thus, it includes data gathered by observing data subjects' behavior (e.g. raw data processed by smart meters, activity logs, or website history). However, it should exclude 'inferred data' and 'derived data', which include personal data that are created by a service provider (e.g. algorithmic results). Different to observed or gathered data, **inferred data are created by the service itself, based on the observed data, not provided by the data subject.**[75] Therefore, the right to data portability **does not include data inferred by a machine-learning process.**

Checklist: data portability[76]

Preparing for requests for data portability

 The controllers know how to recognize a request for data portability and understand when

[74] Article 29 Working Party (2015) Guidelines on the right to data portability. European Commission, Brussels. Available at: http://ec.europa.eu/newsroom/document.cfm?doc_id=45685 (accessed 28 May 2020).

[75] Article 29 Working Party (2015) Guidelines on the right to data portability. European Commission, Brussels, p.8. Available at: http://ec.europa.eu/newsroom/document.cfm?doc_id=45685 (accessed 28 May 2020).

[76] ICO (no date) Right to data portability. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-data-portability/ (accessed 28 May 2020).

the right applies.

- The controllers take into account the requirement for data portability from the earliest stages of conception and design of the AI processing.
- The controllers have a policy for how to record requests they receive verbally.
- The controllers understand when they can refuse a request and are aware of the information they need to provide to individuals if they proceed with such refusal.

Complying with requests for data portability

- The controllers can transmit personal data in structured, commonly used and machine-readable formats.
- The controllers inform users in advance when it is not technically possible to exercise the right of portability by means of a protocol.
- The controllers use secure methods to transmit personal data.
- The controllers have processes to ensure that they respond to a request for data portability without undue delay and within one month of receipt.
- The controllers are aware of the circumstances under which they can extend the time limit to respond to a request.

Additional information

Article 29 Working Party (2016) Guidelines on the right to data portability. European Commission, Brussels. Available at: https://ec.europa.eu/information_society/newsroom/image/document/2016-51/wp242_en_40852.pdf

EBF (2017) European Banking Federation's comments to the Working Party 29 guidelines on the right to data portability. European Banking Federation, Brussels, p.4. Available at: www.ebf.eu/wp-content/uploads/2017/04/EBF_025448E-EBF-Comments-to-the-WP-29-Guidelines_Right-of-data-portabi.._.pdf

ICO (no date) Right to data portability. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-data-portability/

Wallace, N. and Castro, D. (2018) The impact of the EU's new data protection regulation on AI. Center for Data Innovation, Washington, DC / Brussels / London. Available at: www2.datainnovation.org/2018-impact-gdpr-ai.pdf

d) Right to rectification

The right to correct inaccurate data is particularly important in the case of AI, since machine learning algorithms often infer data. These data might affect the data subject, especially if they are produced in advanced steps of the AI life cycle. Inaccurate data inferred during the training phase is not as worrying as in the final phases. Since the

purpose of training data is to train models based on general patterns in large datasets, so individual inaccuracies are less likely to have any direct effect on a data subject.[77] For example, if personal data used to provide information to customers is not correct, such as an erroneous phone number in a dataset, the data subject might suffer more serious harm than if an inferred phone number is used to train a model. However, this most certainly does not mean that the right to rectification does not apply at this stage.

Some concrete types of algorithms, such as Support Vector Machines (SVMs), use some key examples from the training data in order to help distinguish between new examples during deployment. If the data subject requests rectification or erasure of any of this data, the above would not be possible to achieve without having to retrain the model with the rectified data or without deleting the model altogether.[78] This does not, however, render the right to rectification inapplicable.

It is particularly important to keep in mind that if the controller finds that, contrary to the views of the data subject, the data is not inaccurate with regard to the purposes of processing, the controller **does not have to rectify the data.**[79] However, the burden of the proof is placed in the controllers' shoulders. They must provide a good reason to deny rectification, and it is hard to conclude that the damage this could bring to the AI system might serve as a convincing reason. The EDPS has criticized systems that do not include the option to have a set of individual personal data rectified without creating considerable harm to the whole system.[80] In any case, if the controller opts to deny the data subjects' request, they must reply to the data subject with a justified reason for not rectifying the data and, if they wish to, the data subject can then refer the matter to the supervisory authority.[81]

Checklist: right to rectification[82]

[77] Binns, R. (2019) Enabling access, erasure, and rectification rights in AI systems. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/ (accessed 15 May 2020).

[78] Ibid.

[79] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.27. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 28 May 2020).

[80] EDPS (2014) Guidelines on the rights of individuals with regard to the processing of personal data. European Data Protection Supervisor, Brussels, p.18. Available at: https://edps.europa.eu/sites/edp/files/publication/14-02-25_gl_ds_rights_en.pdf (accessed 10 May 2020).

[81] Office of the Data Protection Ombudsman (no date) Right to Rectification. Office of the Data Protection Ombudsman, Helsinki. Available at: https://tietosuoja.fi/en/right-to-rectification (accessed 28 May 2020).

[82] ICO (no date) Right to rectification. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-rectification/ (accessed 28 May 2020).

Preparing for requests for rectification

 The controllers know how to recognize a request for rectification and understand when this right applies.

 The controllers have a policy for how to record requests they receive verbally.

 The controllers understand when they can refuse a request, and are aware of the information they need to provide to individuals when asked to do so.

Complying with requests for rectification

 The controllers are prepared to address the right of rectification of data subjects' data, especially those generated by the inferences and profiles made by the AI solution.

 The controllers have processes in place to ensure that they respond to a request for rectification without undue delay and within one month of receipt.

 The controllers are aware of the circumstances when they can extend the time limit to respond to a request.

 The controllers have appropriate systems to rectify or complete information, or provide a supplementary statement.

 The controllers have procedures in place to inform any recipients if they rectify any data they have shared with them.

Additional information

Binns, R. (2019) Enabling access, erasure, and rectification rights in AI systems. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/

ICO (no date) Right to rectification. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-rectification/

e) Right to erasure

Data subjects have a permanent right to ask the controller for the deletion of their personal data. This might be extremely complicated in some cases, however.[83] Indeed, one must keep in mind that sometimes it may be impossible to fulfil the legal aims of the right to erasure – also known as the right to be forgotten – in AI environments, since the obscurity of the processing might hide some personal data event to the processor (see "y

Understanding transparency and opacity" within this Part III on AI).

[83] Fosch-Villaronga, E., Kieseberg, P. and Li, T. (2018) 'Humans forget, machines remember: artificial intelligence and the right to be forgotten', *Computer Law & Security Review* 34(2): 304-313.

However, the main problem with the right to erasure is that **it might ruin a whole AI system trained on the basis of the data that a subject is asking to erase.** Put simply, algorithms need to retain the data they used for their training. If these data are erased, it could make algorithms less accurate or even break down entirely. Thus, controllers should keep in mind that amending a database that is seriously affected by data erasure might be impossible.

Controllers could consider this to be unacceptable, but the fact is that GDPR do not include any exception to the right to erasure on the basis of the damage caused to a database containing personal data. Some authors, such as Humerick have suggested that "rather than requiring a complete erasure of personal data, controllers and processors should be able to retain information up to the point of erasure. In this way, the AI's machine learning would remain at the point where it progressed, rather than creating forced amnesia." According to him, this could serve well to protect the interests of the data subjects without causing the AI to data-regress. However, it is not easy to be sure that this solution complies with the requirements of the GDPR.

Any recommendation should always focus on the first steps of the life cycle of the product. Technically, it is hard to find secure solutions to the dilemmas posed by the right to erasure once a database has been created. Therefore, controllers should always try to arrive at a simple conclusion: **the best way to avoid catastrophic damage is to prepare for a possible loss of data from the very beginning.**

Finally, the controller must always keep in mind the restrictions to the right to erasure introduced by Article 17(3) of the GDPR. Moreover, national authorities might pose additional restrictions that must be considered.

Checklist: right to erasure[84]

Preparing for requests for erasure

- The controllers know how to recognize a request for erasure and they understand when the right applies.
- The controllers have a policy for how to record requests they receive verbally.
- The controllers understand when they can refuse a request and are aware of the information they need to provide to individuals when doing so.

Complying with requests for erasure

- The controllers have processes in place to ensure that they respond to a request for erasure without undue delay and within one month of receipt.
- The controllers are aware of the circumstances under which they can extend the time limit to respond to a request.

[84] ICO (no date) Right to erasure. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-erasure/ (accessed 28 May 2020).

- The controllers understand that there is a particular emphasis on the right to erasure if the request relates to data collected from children.
- The controllers have procedures to inform any recipients if they erase any data they shared with them.
- The controllers have appropriate methods to erase information.

Additional information

An interview with Tiffany Li on the right to erasure and AI can be found here: www.youtube.com/watch?v=Sozg6yJJkHk

Binns, R. (2019) Enabling access, erasure, and rectification rights in AI systems. ICO blog, 15 October. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/

Fosch-Villaronga, E., Kieseberg, P. and Li, T. (2018) 'Humans forget, machines remember: artificial intelligence and the right to be forgotten', *Computer Law & Security Review* 34(2): 304-313.

Humerick, M. (2018) Taking AI personally: how the E.U. must learn to balance the interests of personal data privacy & artificial intelligence, 34 Santa Clara High Tech. L.J.393. Available at: https://digitalcommons.law.scu.edu/chtlj/vol34/iss4/3

ICO (no date) Right to erasure. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-erasure/

Wallace, N. and Castro, D. (2018) The impact of the EU's new data protection regulation on AI. Center for Data Innovation, Washington, DC / Brussels / London. Available at: www2.datainnovation.org/2018-impact-gdpr-ai.pdf

e) Right to object

Data subjects have the right to object to the processing of their personal data when the controller processes them on the basis of a legitimate interest, or for a task in the public interest. This does not apply to cases where the legal ground for processing was informed consent, since in those cases data subjects could simply withdraw their consent and the controller could no longer process their data. Once data subjects make their request, controllers must cease to process the data, unless they can prove they have compelling and justifiable grounds for continuing to do so, and that these grounds outweigh the data subjects' interests, rights and freedoms.[85]

[85] Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo, p.29. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf (accessed 28 May 2020).

Once the controllers receive an objection to the processing of personal data, and provided that no grounds to refuse apply, **they must stop processing the data immediately**. This may mean that they have to erase stored personal data, as the broad definition of processing under the GDPR includes storing data.

Checklist: right to object

Preparing for objections to processing

 The controllers know how to recognize an objection and they understand when the right applies.

 The controllers have a policy for how to record objections they receive verbally.

 The controllers understand when they can refuse an objection and are aware of the information they need to provide to individuals when doing so.

 The controllers have clear information in their privacy notice about individuals' right to object, which is presented separately from other information on their rights.

 The controllers understand when they need to inform individuals of their right to object, in addition to including it in their privacy notice.

Complying with requests which object to processing

 The controllers have processes in place to ensure that they respond to an objection without undue delay and within one month of receipt.

 The controllers are aware of the circumstances when they can extend the time limit to respond to an objection.

 The controllers have appropriate methods in place to erase, suppress or otherwise cease processing personal data.

Additional information

EDPS (2020) A preliminary opinion on data protection and scientific research. European Data Protection Supervisor, Brussels. Available at: https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf

ICO (no date) The right to object to use your data. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/your-data-matters/the-right-to-object-to-the-use-of-your-data/

Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf

4 Transparency

"This requirement is closely linked with the principle of explicability and encompasses transparency of elements relevant to an AI system: the data, the system and the business models."

- *High-Level Expert Group on AI*[86]

4.1 Ethical and legal issues

This requirement embeds three main different principles: traceability, explainability and communication.[87]

Traceability

The datasets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system. This enables identification of the reasons why an AI-decision was erroneous which, in turn, could help prevent future mistakes. Traceability facilitates auditability as well as explainability.

Explainability

This concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organizational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).

Communication

AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems

[86] High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI, p.18. Brussels, European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 15 May 2020).

[87] Ibid.

must be identifiable as such. In addition, the option to decide against this interaction in favor of human interaction should be provided where needed to ensure compliance with fundamental rights. Beyond this, the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations.

4.2 GDPR provisions: Transparency

4.2.1 Understanding transparency and opacity

In general, transparency means that data subjects are provided with clear information about data processing (see "Transparency" in the "Lawfulness, fairness and transparency" subsection of the "Principles" within Part II of these Guidelines). They must be informed about how and for which purposes their information (including both observed and inferred data about them) is used, no matter whether this information is collected from the data subjects themselves or by others.[88] Data subjects should always be aware of how and why an AI-assisted decision about them was made, or where their personal data was used to train and test an AI system. Controllers must keep in mind that in such cases, transparency is even more important than when they have no direct relationship with the data subjects.

In general, transparency must be guaranteed by using a number of complementary tools. Naming a DPO, who then serves as a single point of contact for queries from data subjects, is an excellent option. Preparing adequate records of processing for the supervisory authorities, or performing DPIAs, are also highly recommended measures to promote transparency. And undertaking analysis that evaluate the effectiveness and accessibility of the information provided to the data subjects helps to ensure the efficient implementation of this principle.

The main challenge with AI is that it encompasses a range of techniques that are very different from each other. Some are very simple, so it is easy for the controller to provide all the necessary information. Others, such as deep learning, suffer from serious problems in terms of transparency. This is often referred to as the 'black box' issue, which introduces the opacity issue in the AI framework, a circumstance that renders transparency difficult to achieve. Indeed, opacity is one of the main threats against fair AI, since it directly defies the need for transparency. There are at least three types of opacity that are inherent in AI to a greater or lesser extent: (1) as intentional corporate or state secrecy; (2) as technical illiteracy; and (3) epistemic opacity.

4.2.1.1 Opacity as intentional corporate or state secrecy

This kind of opacity can be legitimate under the protection of industrial secret regulations. It can also respond to legitimate interests, such as preserving competitive advantages, preserving the security of the system, or preventing malign users from

[88] Articles 13 and 14 of the GDPR.

gaming the system. However, it should be compatible with the incorporation of independent certification systems that are capable of accrediting that the mechanism meets the requirements of the GDPR. In most cases, furnishing the data subject with the information they need to protect their interests, without at the same time disclosing trade secrets, should not be problematic.[89] This is due to the simple fact that subjects do not need to understand in detail how the system works, only how it might cause damage to their interests, rights and freedoms.

4.2.1.2 Opacity as technical illiteracy

This kind of opacity derives from the specific skills required to design and program the algorithms, and the ability to read and write the code. We could say that the codes used in AI are a mystery for the vast majority of the population, who lack this specific knowledge, but this should not be an impediment to the fulfilment of the obligation of information stipulated by the GDPR. The ability to understand computer language must not be a barrier for providing an understandable explanation of the purpose of an AI system, not only to the stakeholders who are subject to profiling or automated decision, but to everyone else.

4.2.1.3 Epistemic opacity

This opacity arises from the characteristics of machine-learning algorithms and the scale required to apply them usefully. It is related to the fact that certain algorithmic models are not interpretable by humans. Put simply, the transit between the inputs that the model receives and the outputs that it throws out is inscrutable in terms of human understanding. At the regulatory level, there is no ban on the use of this type of model, although it is advisable to follow the precautionary principle when using it, since the lack of interpretability could aggravate the difficulties of identifying biases in the model, which could in turn lead to discriminatory results, or false or spurious correlations. Of course, not all machine-learning models are opaque in this sense.

4.2.1.4 The preference for transparent tools

In general, controllers should always provide for the development of more understandable algorithms over less understandable ones. Trade-offs between the **explainability, transparency and best performance of the system** must be appropriately balanced based on the context of use. For instance, in healthcare, the accuracy and performance of the system may be more important that its explainability; in policing, explainability is much more crucial to justify behavior and outcomes of law enforcement. In other areas, such as recruitment, both accuracy and explainability are similarly valued.[90] If a service can be offered through both by an easy to understand and

---

[89] Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf (accessed 20 May 2020).

[90] SHERPA (2019) Guidelines for the ethical use of AI and big data systems. SHERPA project, p.26. Available at: www.project-sherpa.eu/wp-content/uploads/2019/12/use-final.pdf (accessed 15 May 2020).

an opaque algorithm – that is, when there is no trade-off between explainability and performance – the controller should opt for the one that is more interpretable.

If controllers have no choice but to use an opaque model, they should at least try to find technical solutions to the lack of interpretability. Of course, the sense to which an extension in explainability is achieved is extremely hard to measure precisely. For more information, see the section on "Right not be subject to automated decision-making" within Part II section "Data subject's rights" of these Guidelines'. If explanations are hard to find for controllers, they should seek external advice. The possibility of using independent audits may again be a reasonable option.

Additional information

EDPS (2015) Opinion 7/2015. Meeting the challenges of big data: A call for transparency, user control, data protection by design and accountability. European Data Protection Supervisor, Brussels. Available at: https://edps.europa.eu/sites/edp/files/publication/15-11-19_big_data_en.pdf

High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI. European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

ICO (2020) Explaining decisions made with AI. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf

Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf

SHERPA (2019) Guidelines for the ethical use of AI and big data systems. Sherpa project. Available at: https://www.project-sherpa.eu/wp-content/uploads/2019/12/use-final.pdfwww.project-sherpa.eu/wp-content/uploads/2019/12/use-final.pdf

5 Fairness, diversity and non-discrimination

"In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment. This requirement is closely linked with the principle of fairness."

- *High-Level Expert Group on AI*[91]

5.1 Ethical principles

Avoidance of unfair bias

Datasets used by AI systems, both for training and operation, may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalization. Harm can also result from the intentional exploitation of (consumer) biases, or by engaging in unfair competition, such as the homogenization of prices by means of collusion or a non-transparent market.

Identifiable and discriminatory bias should be removed in the collection phase where possible. The way in which AI systems are developed (e.g. algorithms' programming) may also suffer from unfair bias. This can be counteracted by putting in place oversight processes to analyze and address the system's purpose, constraints, requirements and decisions in a clear and transparent manner. Moreover, hiring developers from diverse backgrounds, cultures and disciplines can ensure a diversity of opinions – and should therefore be encouraged.

Accessibility and universal design

Systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for minors or persons with disabilities, which are present in all societal groups, is of particular importance. AI systems should not have a one-size-fits-all approach and should consider Universal Design principles, addressing the widest possible range of users, following relevant accessibility standards. This will enable equitable access and active participation of all people in existing and emerging computer-mediated human activities, and with regard to assistive technologies.

Stakeholder participation

To develop AI systems that are trustworthy, it is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle. It is beneficial to solicit regular feedback even after deployment, and set up longer-term mechanisms for stakeholder participation, for example by ensuring workers' information, consultation and participation throughout the whole process of implementing AI systems at organizations.

[91] High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI, p. 15 and ff. European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 15 May 2020).

5.2 GDPR provisions

According to Recital 71 of the GDPR, "the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organizational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimized, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect."

This paragraph is strictly linked to **Article 21 of the EU Charter of Fundamental Rights**, which states that "[a]ny discrimination based on any ground such as sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited." Meanwhile, the EDPB provides a definition of fairness in its Guidelines on Data Protection by Design and by Default, which states that "[f]airness is an overarching principle which requires that personal data shall not be processed in a way that is detrimental, discriminatory, unexpected or misleading to the data subject."[92]

Discrimination is, therefore, a dramatic violation of the fairness principle. However, in the AI field, biases constitute a formidable threat against this principle, because they could lead to potential stigmatization or discrimination of isolated individuals or entire communities.[93]

5.2.1.1 Bias: the causes

Biases can be caused by a number of different issues, and when data is gathered, it may contain **socially constructed biases, inaccuracies, errors and mistakes.** There are multiple reasons that explain these biases. Sometimes, it might happen that datasets are biased due to **malicious actions.** Feeding malicious data into an AI system may change its behavior, particularly with self-learning systems.[94] For instance, in the case of the chatbot Tay, developed by Microsoft, a huge number of internet users started posting racist and sexist comments that served to feed the algorithm. As a result, Tay started sending racist and sexist tweets after just a few hours of operation. In other cases, **data are simply of poor quality** and this creates bias. For example, data taken from social media platform present serious risks for researchers, due to the characteristics of the

[92] EDPB (2019) Guidelines 4/2019 on Article 25 Data Protection by Design and by Default (version for public consultation). European Data Protection Board, Brussels. Available at: https://edpb.europa.eu/our-work-tools/public-consultations-art-704/2019/guidelines-42019-article-25-data-protection-design_es (accessed 20 May 2020).

[93] Mittelstadt, B. and L. Floridi, L. (2016) 'The ethics of big data: current and foreseeable issues in biomedical context', *Science and Engineering Ethics* 22(2): 303-341.

[94] High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI. European Commission, Brussels, p.17. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 20 May 2020).

online environment, which does not guarantee the accuracy and representativeness of the data.

Another reason for biases is **imbalanced training data** (see Box 8), which arises when the proportion of different categories in the training data is not balanced. For instance, in the context of clinical trials, there might be much more data from males than females. In such cases, females are likely to be discriminated against by the resulting AI model. Therefore, issues related to the composition of the databases used for training raise crucial ethical and legal issues, not only issues related to efficiency or of a technical nature.

Box 8. Biases caused by imbalanced data training

The Beauty.AI case

Launched in 2016, the Beauty.AI tool was created to select "the First Beauty Queen or King Judged by Robots", using age and facial recognition algorithms. Seven thousand people sent in their pictures through an app, but most of the 44 winners were white; only a handful were Asian, and only one had dark skin. This was despite the fact that, although the majority of contestants were white, many people of color submitted photos, including large groups from Africa and India. This was immediately considered a racist result, due to poor selection of the training dataset. The main problem was that the data the project used to establish standards of beauty were mainly composed by white people. Although the developers did not build the algorithm to treat light skin as a sign of beauty, the input data effectively led the robot judges to reach that conclusion.[95]

The Amazon recruiting tool

In December 2018, Amazon scrapped its AI recruiting tool when the company discovered that the AI system showed bias against women. Amazon had been building computer programs since 2014 to review job applicants' resumes, with the aim of mechanizing the search for top talent. The tool used AI to score job candidates from one to five stars. In 2015, however, Amazon discovered that the tool was not rating candidates for software developer jobs and other technical posts in a gender-neutral way. This was because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.[96]

Thirdly, the training data may reflect **past discrimination produced by societal trends** (see Box 9). If controllers use historical data, they should be aware of the probable

[95] LEVIN, S. (2016) 'A BEAUTY CONTEST WAS JUDGED BY AI AND THE ROBOTS DIDN'T LIKE DARK SKIN', *THE GUARDIAN*, 8 SEPTEMBER. AVAILABLE AT: WWW.THEGUARDIAN.COM/TECHNOLOGY/2016/SEP/08/ARTIFICIAL-INTELLIGENCE-BEAUTY-CONTEST-DOESNT-LIKE-BLACK-PEOPLE (ACCESSED 26 MAY 2020).

[96] Dastin, J. (2018) 'Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters*, 10 October. At: www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

differences between social contexts compared to the present day. Otherwise, biases will be unavoidable. Sometimes the biases come from the different social contexts of the community that provided the data and the community that is meant to use the algorithm. If the controller does not pay careful attention to this, again biases will probably be present in the tool.

Box 10. Biases produced by societal trends

In the past, loan applications from women were rejected more frequently than those from men, due to prejudice. In this case, any AI model trained on historical data is likely to reproduce the same pattern of discrimination. These issues can occur even if the training data does not contain any protected characteristics, such as gender or race. A variety of features in the training data are often closely correlated with protected characteristics (e.g. occupation, race, etc.). These 'proxy variables' enable the model to reproduce patterns of discrimination associated with those characteristics, even if its designers did not intend this.

These problems can occur in any statistical model. However, they are more likely to occur in AI systems because they can include a greater number of features, and may identify complex combinations of features that are proxies for protected characteristics. Many modern machine-learning methods are more powerful than traditional statistical approaches because they are better at uncovering non-linear patterns in high dimensional data. However, these also include patterns that reflect discrimination.[97]

Finally, it is possible that biases are caused **by a poorly designed AI tool** (see Box 11). It might happen that the designer introduces correlations by proxies that do not work well with reality. If this is the case, the model will make inaccurate predictions, since its conceptual basis are not solid.

Box 11. Bias caused by a poorly designed AI tool: algorithms

The US healthcare system uses commercial algorithms to guide health decisions. Obermeyer et al.[98] found evidence of racial bias in one widely used algorithm, which meant that, among black and white patients assigned the same level of risk by the algorithm, the black patients were sicker than the white ones. The authors estimated that this racial bias reduced the number of black patients identified for extra care by more than half. Bias occurred because the algorithm used health costs as a proxy for health needs. Less money was spent on black patients with the same level of need as white patients, and the algorithm thus falsely concluded that black patients were healthier than equally sick

[97] ICO (2020) AI auditing framework: draft guidance for consultation, p.54. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (accessed 26 May 2020).

[98] Obermeyer, Z. et al. (2019) 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science,* 25 October, 447-453.

white patients. In reality, the minor expenditure was caused by a number of racially biased factors, such as different access to treatment, levels of trust in the system, imbalances caused by healthcare givers, etc.

5.2.1.2 Resolving biases

There are different strategies that could help to avoid biases or to correct them. While creating the databases that will serve to build an AI model, controllers **should make every effort to avoid unbalanced or mistaken data.** Identifiable and discriminatory bias should be removed in the dataset-building phase where possible.[99]

If the origin of the bias is related to the training dataset, the controller should search for an **adequate selection of data to be used in the training phase,** to avoid the results of the subsequent model being incorrect or discriminatory.[100] An AI model must "be trained **using relevant and correct data and it must learn which data to emphasize.** The model must not emphasize information relating to racial or ethnic origin, political opinion, religion or belief, trade union membership, genetic status, health status or sexual orientation if this would lead to arbitrary discriminatory treatment (emphasis added)."[101]

Furthermore, people with disabilities should be included when sourcing data to build models, and in testing, to create a more inclusive and robust system. If this process is performed adequately, the bias will probably vanish. For example, in the racial bias in health algorithms case study (see Box 11), it was possible to reformulate the algorithm (in this case, so that it no longer used costs as a proxy for needs) and eliminate racial bias in predicting who needed extra care. Indeed, changing the indicator for health, from predicted costs to the number of chronic medical conditions, increased the percentage of black patients receiving better healthcare from 17% to 46%. This is an excellent example of increasing fairness by reformulating an algorithm.

However, controllers should always keep in mind that what makes fighting biases so particularly complex is that selecting a dataset involves making decisions and choices – which may, at times, almost be done **unconsciously**. By contrast, coding a traditional, deterministic algorithm is always a deliberate operation. Indeed, humans are always the intelligence behind a development - even when it is embedded in algorithms that we think are neutral. Whoever builds a dataset is, to some extent, building it in their own

[99] Recital 71 of the GDPR.

[100] For a definition of direct and indirect discrimination, see, for instance, Article 2 of Council Directive 2000/78/EC of 27 November 2000, which establishes a general framework for equal treatment in employment and occupation. See also Article 21 of the Charter of Fundamental Rights of the EU.

[101] Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo, p.16. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf (accessed 15 May 2020).

image, to reflect their own worldview, values or, at the very least, the values which are more or less inherent in the data gathered from the past.[102]

In light of this, it is important that the teams in charge of selecting the data to be integrated into a dataset should comprise **people that reflect the diversity that the AI development is expected to show.** At present, this is a major challenge. In terms of gender, for example, women comprise only 15% of AI research staff at Facebook and 10% at Google, and there is no public data on trans workers or other gender minorities. In terms of race, the gap is even starker: only 2.5% of Google's workforce is black, while Facebook and Microsoft are each at 4%.[103] Controllers should make every effort to ensure that their **teams better reflect diversity and implement accurate data that reflects this.**

In summary, algorithms' development processes **should always include a careful monitoring of possible biases.** Internal and external reviews should pay special attention to this issue. Datasets built for validation purposes should be carefully selected to ensure an adequate incorporation of data pertaining to subjects from different sectors of society, in terms of age, race, gender, disabilities, etc. Fortunately, there are a lot of technical tools devoted to eradicating biases in AI models.[104] The IEEE P7003TM Standard for Algorithmic Bias Considerations is particularly interesting at the moment.[105]

However, none of them offers a magical solution, or 'silver bullet', applicable to all types of algorithms. In most cases, the right solution will depend on the multiple variables involved in the algorithm. Controllers should aim to eradicate biases as far as possible, and be honest about the final results of their efforts. If biases are uncovered, the AI solution should be trained again. If **unfair biases cannot be erased from the model, its deployment should not proceed.**

Checklist: bias

- The controller has established a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data and for the algorithm design.
- The controller assesses and acknowledges the possible limitations stemming from the composition of the used datasets.

[102] CNIL (2017) How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence. Commission Nationale de l'Informatique et des Libertés, Paris, p.34. Available at: www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf (accessed 15 May 2020).

[103] West, S.M., Whittaker, M. and Crawford, K. (2019) Discriminating systems: gender, race and power in AI. AI Now Institute, New York, p.3. Available at: https://ainowinstitute.org/discriminatingsystems.html (accessed 15 May 2020).

[104] ICO (2020) AI auditing framework: draft guidance for consultation. Information Commissioner's Office, Wilmslow, p.55-56. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (accessed 15 May 2020).

[105] See: https://ethicsinaction.ieee.org/ (accessed 17 May 2020).

- The controller has considered the diversity and representativeness of the data used.
- The controller has tested for specific populations or problematic use cases.
- The controllers used the available technical tools to improve their understanding of the data, model and performance.
- The controller has put in place processes to test and monitor for potential biases during the development, deployment and use phases of the AI system.
- The controller has implemented a mechanism that allows others to flag issues related to bias, discrimination or poor performance of the AI system.
- The controller has established clear steps and ways of communicating on how and to whom such issues can be raised.
- The controller has considered others, potentially indirectly affected by the AI system, in addition to the (end-)users.
- The controller has assessed whether there is any possible decision variability that can occur under the same conditions.
- In case of variability, the controller has established a measurement or assessment mechanism of the potential impact of such variability on fundamental rights.
- The controller has implemented a quantitative analysis or metrics to measure and test the applied definition of fairness.
- The controller has established mechanisms to ensure fairness in the AI systems, and has considered other potential mechanisms.

Additional information

CNIL (2017) How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence. Commission Nationale de l'Informatique et des Libertés, Paris. Available at: www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf

EDPB (2019) Guidelines 4/2019 on Article 25 Data Protection by Design and by Default (version for public consultation). European Data Protection Board, Brussels. Available at: https://edpb.europa.eu/our-work-tools/public-consultations-art-704/2019/guidelines-42019-article-25-data-protection-design_es

ICO (2020) AI auditing framework: draft guidance for consultation. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

Mittelstadt, B. and Floridi, L. (2016) 'The ethics of big data: current and foreseeable issues in biomedical context', *Science and Engineering Ethics* 22(2): 303-341.

Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at:

https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf

West, S.M., Whittaker, M. and Crawford, K. (2019) Discriminating systems: gender, race and power in AI. AI Now Institute, New York, p.3. Available at: https://ainowinstitute.org/discriminatingsystems.html

6 Societal and environmental well-being

"In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, such as for instance the Sustainable Development Goals. Ideally, AI systems should be used to benefit all human beings, including future generations."

- *High-Level Expert Group on AI*[106]

6.1 Ethical principles

Sustainable and environmentally friendly AI

AI systems promise to help tackle some of our most pressing societal concerns, but this must be achieved in the most environmentally friendly way possible. The system's development, deployment and use processes, as well as its entire supply chain, should be assessed in this regard. This includes measures such as a critical examination of its resource use and energy consumption during training, and opting for less environmentally harmful choices where available. Measures securing the environmental friendliness of AI systems' entire supply chain should also be encouraged.

Social impact

Ubiquitous exposure to social AI systems in all areas of our lives - be it education, work, care or entertainment - may alter our conception of social agency, or impact our social relationships and attachment. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could also affect people's physical and mental wellbeing. The effects of these systems must therefore be carefully monitored and considered.

[106] High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI, p.19. European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 15 May 2020).

Society and democracy

Beyond assessing the impact of an AI system's development, deployment and use on individuals, this impact should also be assessed from a societal perspective, taking into account its effect on institutions, democracy and society at large. The use of AI systems should always be given careful consideration, particularly in situations relating to democratic processes, including not only political decision-making but also electoral contexts.

6.2 GDPR provisions: legitimacy

The GDPR does not include specific provisions related to societal and environmental well-being. However, Article 5(1)(b) states that "personal data shall be collected for specific, explicit and legitimate purposes". Through this clause, the GDPR introduces the concept of legitimacy in the data protection context.

However, legitimacy is a fuzzy concept that is not at all defined by the GDPR (see "Lawfulness, fairness and transparency principle" within Part II section "Principles" of these Guidelines). The Article 29 Working Party states that it "means that the purposes must be 'in accordance with the law' in the broadest sense. This includes all forms of written and common law, primary and secondary legislation, municipal decrees, judicial precedents, constitutional principles, fundamental rights, other legal principles, as well as jurisprudence, as such 'law' would be interpreted and taken into account by competent courts." Therefore, it must be understood as a very broad concept that embeds social and environmental well-being considerations.

In the 'White paper on artificial intelligence: a European approach to excellence and trust', the authors note that "[g]iven the increasing importance of AI, the environmental impact of AI systems needs to be duly considered throughout their lifecycle and across the entire supply chain, e.g. as regards resource usage for the training of algorithms and the storage of data".[107]

Further concrete recommendations for AI development that are oriented to societal and environmental well-being can be found in the 'Report from the Commission to the European Parliament, the Council and the European economic and social committee: report on the safety and liability implications of artificial intelligence, the internet of things and robotics'.[108] This kind of ethical recommendations should be carefully

[107] European Commission (2020) White Paper on artificial intelligence: a European approach to excellence and trust. European Commission, Brussels, p.3. Available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (accessed 26 May 2020).

[108] European Commission (2020) Report from the Commission to the European Parliament, the Council and the European economic and social committee: report on the safety and liability implications of artificial intelligence, the internet of things and robotics. European Commission, Brussels. Available at: https://ec.europa.eu/info/sites/info/files/report-safety-liability-artificial-intelligence-feb2020_en_1.pdf (accessed 26 May 2020).

considered by AI developers before processing personal data, since they are clearly linked to their legitimacy.

7 Accountability

"The requirement of accountability complements the above requirements, and is closely linked to the principle of fairness. It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use."

- *High-Level Expert Group on AI*[109]

7.1 Ethical principles

Auditability

Auditability means enabling the assessment of algorithms, data and design processes. Evaluation by internal and external auditors, and the availability of such evaluation reports, can contribute to the trustworthiness of the technology. In applications affecting fundamental rights, including safety-critical applications, AI systems should be open to independent auditing. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available, however.

Minimization and reporting of negative impacts

It is essential to ensure both the ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome. Identifying, assessing, documenting and minimizing the potential negative impacts of AI systems is especially crucial for those (in)directly affected. Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system. The use of impact assessments (e.g. red teaming or forms of algorithmic impact assessment), both prior to and during the development, deployment and use of AI systems, can help to minimize negative impacts. These assessments must be proportionate to the risk that the AI systems pose.

Trade-offs

When implementing the above requirements, tensions may arise between them, which may lead to inevitable trade-offs. Such trade-offs should be addressed in a rational and methodological manner within the state of the art. This entails that relevant interests and values implicated by the AI system should be identified and that, if conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights. In situations in which no ethically acceptable trade-offs can be identified, the development, deployment and use of the AI

[109] High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI, p.19. European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 15 May 2020).

system should not proceed in that form. Any decision about which trade-off to make should be explained and properly documented. The decision-maker must be accountable for the manner in which the appropriate trade-off is made, and should continually review the appropriateness of the resulting decision to ensure that necessary changes can be made to the system where needed.

Redress

When unjust adverse impact occurs, accessible mechanisms should be in place that ensure adequate redress. Knowing that redress is possible when things go wrong is key to ensuring trust. Particular attention should be paid to vulnerable persons or groups.

7.2 GDPR provisions

7.2.1 Accountability

According to Article 5(2) of the GDPR, the controller shall be responsible for, and must be able to demonstrate, compliance with all principles of the GDPR mentioned at Article 5(1). This includes the principle of accountability see "Accountability principle" within Part II section "Principles" of these Guidelines).

The accountability principle in the GDPR is risk-based: the higher the risk of data processing to the fundamental rights and freedoms of data subjects, the greater the measures needed to mitigate those risks.[110] The accountability principle is based on several compliance duties for data controllers, including: transparency duties (Articles 12-14); guaranteeing the exercise of data protection rights (Articles 15-22); keeping records of the data-processing operations (Article 30); notifying eventual data breaches to a national supervisory authority (Articles 33) and to the data subjects (Article 34); and, in cases of higher risk, hiring a DPO and carrying out a DPIA (Article 35).

Since the processing of personal data in AI systems might often be considered as high risk,[111] the developer of AI will often need to have a DPO and perform a DPIA. The next two sections address these two specific accountability duties.

7.2.2 Risk assessment and DPIAs

A DPIA is a process in which the data controller, before starting a data-processing procedure with high risk to the fundamental rights and freedoms of data subjects, assesses the impact of the envisaged processing operations on the protection of personal data (Article 35(1)).

Determining if the data processing is of high risk is not an easy task, however. Article 35(3) lists three cases: (1) a systematic and extensive evaluation of personal aspects

---

[110] See Articles 24, 25 and 32 of the GDPR, which require controllers to take into account the "risks of varying likelihood and severity for the rights and freedoms of natural persons" when adopting specific data protection measures.

[111] See, in particular, Article 35(3)(a), according to which data processing is considered as high risk in cases of, inter alia, "a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person".

relating to natural persons, which is based on automated processing, including profiling and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person; (2) processing on a large scale of special categories of data referred to in Article 9(1), or of personal data relating to criminal convictions and offences referred to in Article 10; and (3) a systematic monitoring of a publicly accessible area on a large scale.

With regard to innovative technologies, the Article 29 Working Party clarified some examples, such as "combining use of fingerprint and face recognition for improved physical access control" and "certain "Internet of Things" applications". These data-processing operations are considered as high risk "because the use of such technology can involve novel forms of data collection and usage, possibly with a high risk to individuals' rights and freedoms. Indeed, the personal and social consequences of the deployment of a new technology may be unknown."[112]

If the processing is high risk, then a DPIA should be conducted following Article 35(7) of the GDPR. Recital 90 of the GDPR further clarifies that the assessment of risk should be done using two parameters: the **likelihood** and **severity** of high risk, taking into account the nature, scope, context and purposes of the processing and the sources of risk. Several national supervisory authorities have issued guidance on how to assess these risks, such as the Agencia Española de Protección de Datos Personales, the Information Commissioner's Office, the Irish Data Protection Commission, the Commission Nationale de l'Informatique et des Libertés, among others (see "DPIA" in Part II, section "Main Tools and Actions" within Part II of these Guidelines).

In certain situations, if the result of the DPIA is that the intended processing activity has a high risk of causing harm to the fundamental rights and freedoms of data subjects, the controller should request the opinion of the national supervisory authority, as prescribed by Article 36 of the GDPR. Some Member States have issued lists that contain examples of data-processing activities that would trigger this mandatory consultation; among those examples, we can identify situations that match with AI techniques and, in some cases, go as far as expressly including AI. Supervisory authorities can require the adoption of certain measures to mitigate the risk, if possible, or forbidding the use of AI if it is not possible.

Checklist: is a DPIA necessary?

ü The controller determined the jurisdictions where data-processing activities will take place.

ü The controller checked if those jurisdictions have enacted lists indicating the processing that require a DPIA and saw if the intended data processing activities that involve AI are covered by those provisions.

ü If the controllers are unsure of the necessity of carrying out a DPIA, they consult with the DPO or, in lieu of, legal department of the controller.

[112] Article 29 Working Party (2017) Guidelines on the Data Protection Impact Assessment, WP248, pp. 10. European Commission, Brussels. Available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236 (accessed 20 May 2020).

ü If necessary, the controller carried out a DPIA.

ü If necessary, the controller filed a prior consultation with the appropriate supervisory authority.

ü If changes were suggested, the controller followed the advice of the supervisory authority.

7.2.3 Processor due diligence

The accountability principle see "Accountability principle" within Part II section "Principles" of these Guidelines) is also present when a controller chooses to require the services of a processor. In this regard, Article 28(1) of the GDPR[113] requires controllers to perform certain due diligence actions, and prior to providing processors with access to the personal data for the performance of data-processing activities. As with other provisions of the GPDR, it is not stated which specific actions a controller should carry out when evaluating processors. The only criteria provided by the GDPR is that **controllers should judge processors on the basis of their ability to demonstrate that they can carry out processing activities in compliance with the GDPR.**

Therefore, a researcher conducting AI development that needs to hire a third party for certain processing activities would need to ask two questions: (1) what type of conduct is expected to demonstrate compliance with this obligation; and (2), if some form of positive action is expected, how should controllers proceed to carry such due diligence?

For the first question, the GPDR indicates that if controllers intend to remain compliant with the GDPR, they can only retain a processor that is able to demonstrate their compliance with the GDPR. Therefore, controllers need to request information to assess this. In other words, the GDPR expects controllers to actively ask their potential processor about this; it is not sufficient to rely on a representations and warranties clause in the data-processing agreement (see "Integrity and confidentiality principle" within Part II section "Principles" of these Guidelines).

As for how controllers should carry out this due diligence, again the GDPR does not provide concrete issues to analyze. Nevertheless, certain national supervisory authorities have proposed topics to consider, such as whether the processor follows industry standards, to request the provision of both legal and technical information about how the processor processes personal data, if they adhere to a code of conduct, or if they have gone through a certification scheme.[114]

Besides these general considerations, and depending on how the processing requested to this third party integrates within the framework of the developed AI, further questions

---

[113] 'Article 28 Processor 1. "Where processing is to be carried out on behalf of a controller, the controller shall use only processors providing sufficient guarantees to implement appropriate technical and organizational measures in such a manner that processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject."

[114] ICO (no date) Guide to the General Data Protection Regulation (GDPR), What responsibilities and liabilities do controllers have when using a processor? Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/contracts-and-liabilities-between-controllers-and-processors-multi/responsibilities-and-liabilities-for-controllers-using-a-processor/ (accessed 20 May 2020).

should be asked. In this regard, **any question that the controllers would ask themselves when developing the AI should be asked to the processor.** We defer to the issues posed in the Checklist for further guidance.

Checklist: processor due diligence

ü The controllers required information regarding where the data-processing activities will take place, and: (1) carry out the case law review suggested below; and (2) assess if the jurisdictions, in case of non-EU countries, are deemed as adequate by the EU Commission.

ü The controllers reviewed case law from the national supervisory authorities where the processor operates to check for potential sanctions.

ü The controllers required proof of adherence to a code of conduct or certification.

ü The controllers required proof of relevant ISO certification.

ü The controllers required a copy of records of processing activities.

ü The controllers enquired about the development process of the AI, in particular which kind of data were used for training the AI and the data that the AI needs to operate and deliver a useful result.

7.2.4 DPOs

DPOs play a crucial role when designing and implementing data-processing activities in a GDPR-compliant manner. They are another safeguard that the GDPR mandates on certain occasions and, in general, it is recommended to appoint such a figure. The Article 29 Working Party considers that this "is a cornerstone of accountability and that appointing a DPO can facilitate compliance".[115]

Article 37(1) of the GDPR[116] outlines when controllers and processors should appoint a DPO. In the case of AI development, and as explained previously**, the appointment of a DPO is (almost) certainly necessary, as many AI systems process personal data, which would make them applicable under the conditions described in Article 37(1)(a) and (b) in most situations.** This opinion is shared by, as an example, the

[115] Article 29 Working Party (2017) Guidelines on Data Protection Officers ('DPOs'), p.4. European Commission, Brussels.

[116] Article 37. Designation of the data protection officer. 1. The controller and the processor shall designate a data protection officer in any case where: (a) the processing is carried out by a public authority or body, except for courts acting in their judicial capacity; (b) the core activities of the controller or the processor consist of processing operations which, by virtue of their nature, their scope and/or their purposes, require regular and systematic monitoring of data subjects on a large scale; or (c) the core activities of the controller or the processor consist of processing on a large scale of special categories of data pursuant to Article 9 and personal data relating to criminal convictions and offences referred to in Article 10.

Spanish supervisory authority.[117] However, neither the Article 29 Working Party nor the EDPB has specifically stated that a DPO is mandatory if a controller or processor engages in data-processing activities that involve AI. Nevertheless, the Article 29 Working Party has pointed out that **profiling activities can be considered as activities that trigger the mandatory appointment of a DPO**[118] if, as pointed out above, these profiling activities involve AI.

It would be useful if each Member States' regulations on the need for DPOs expanded the list of activities that demand the appointment of a DPO or, at least, provided clear examples that could help to interpret which data-processing activities carried out by controllers and processor demand such an appointment.

If a DPO has to be appointed, for any of the reasons mentioned above, it is necessary to have their participation in the DPIA (required by Article 39(1)(c)) as well as any other issue related to data protection within the entity (as prescribed by Article 39(1)(a)). This may include reviewing a potential processor, as described in the previous item. Therefore, the researchers involved in the development of the AI should consult with the DPO regarding the data-protection issues that might arise during the development of the AI. For example, the role of the DPO, in connection to AI systems, is also relevant for collaborating in drafting an appropriate notice, as required by Articles 13 and 14 as it corresponds, to properly communicate data subjects how the AI operates and what consequences it might have on them.

Checklist: DPOs

ü The controllers checked if the institution has already appointed a DPO.

ü If not, they checked with the legal department if the intended data-processing activities trigger the appointment of a DPO, either by looking at European authoritative interpretations, local regulations, local authoritative interpretations, case law - both local and European - and, finally, academic interpretations.

ü The controllers required the appointment of DPOs if necessary, and their involvement in the AI development process as necessary.

ü As a general rule, the DPO should be aware of every step taken to allow room for their intervention if deemed relevant.

Additional information

Agencia Española de Protección de Datos Personales (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción, p.35. Agencia

[117] Agencia Española de Protección de Datos Personales (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción, p.35. Agencia Española de Protección de Datos Personales, Madrid. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 20 May 2020).

[118] Article 29 Working Party (2017) Guidelines on Data Protection Officers ('DPOs'), p.4. European Commission, Brussels.

Española de Protección de Datos Personales, Madrid. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf

Article 29 Working Party (2010) Opinion 3/2010 on the principle of accountability. European Commission, Brussels. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp173_en.pdf

Article 29 Working Party (2017) Guidelines on the Data Protection Impact assessment (DPIA), pp. 9-10. European Commission, Brussels. Available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236

AI: Step by Step Process

Iñigo de Miguel Beriain (UPV/EHU)

Acknowledgements: The author thankfully acknowledges advice, input and feedback on drafts from Andres Chomsky, Oliver Feeney, Gianclaudio Malgieri Aurélie Pols and Marko Sijan. Needless to say, all mistakes are my full responsibility.

This part of The Guidelines has been reviewed and validated by Marko Sijan, Senior Advisor Specialist, (HR DPA)

Introduction part B

This second part of The Guidelines is built on the basis of a step-by-step model, the CRISP-DM model,[119] which is widely employed for explaining the stages included in the development of data analytics and data-intensive AI tools. Indeed, it was the tool selected by the SHERPA project to develop their Guidelines for the Ethical Development of AI and Big Data Systems.[120] These six steps are: business understanding; data understanding; data preparation; modeling; evaluation; and

[119] Shearer, C. (2000) 'The CRISP-DM model: the new blueprint for data mining', *Journal of Data Warehousing* 5(4): 13-23. Available at: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf (accessed 15 May 2020).

[120] SHERPA project (2019) Guidelines for the ethical development of AI and big data systems: an ethics by design approach. SHERPA Project. Available at: www.project-sherpa.eu/wp-content/uploads/2019/12/development-final.pdf (accessed 15 May 2020).

deployment. This is not a fixed classification, since quite often developers mix some of these stages. For instance, a trained algorithm might be improved after the validation stage through a renewed training.

Nevertheless, it must be highlighted that some of the ethical and legal requirements regarding AI development must be evaluated through the life cycle of an AI development on a continuous basis. Controllers must monitor the ethical legitimacy of processing, and its unexpected effects They should also assess the possible collateral impact of such processing in a social environment, beyond the initially conceived limitations of purpose, duration in time and extension.[121] And this must be done all along the life cycle of an AI tool, according to Article 25 of the GDPR. As the Article 29 Working Party stated,

> *"Controllers should carry out frequent assessments on the data sets they process to check for any bias, and develop ways to address any prejudicial elements, including any over-reliance on correlations. Systems that audit algorithms and regular reviews of the accuracy and relevance of automated decision-making including profiling are other useful measures. Controllers should introduce appropriate procedures and measures to prevent errors, inaccuracies or discrimination on the basis of special category data. These measures should be used on a cyclical basis; not only at the design stage, but also continuously, as the profiling is applied to individuals. The outcome of such testing should feed back into the system design."*[122]

An additional idea that deserves a thought is that AI is a common label that encompasses a variety of different technologies. A fundamental distinction must be traced between supervised machine learning (input data labelled by humans is given to an algorithm, which then defines the rules based on examples which are validated cases) and unsupervised learning (unlabelled input data is given to an algorithm, which carries out its own classification and is free to produce its own output when presented with a pattern or variable). Supervised learning requires supervisors to teach the machine the output it must produce, i.e. they must "train" it. In principle, supervised learning is easier to understand and monitor.[123] Moreover, since the datasets used in training processes are selected by the trainers, we might handle some of the most worrying challenges posed by these technologies quite reasonably. Unsupervised AI, instead, and more especially techniques such deep learning, needs a more sophisticated monitor and

---

[121] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.7. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 15 May 2020).

[122] Article 29 Working Party (2017) Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679. Adopted on 3 October 2017 as last Revised and Adopted on 6 February 2018. European Commission, Brussel, p.28. Available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053 (accessed 15 May 2020).

[123] CNIL (2017) How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence. Commission Nationale de l'Informatique et des Libertés, Paris, p.17. Available at: www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf (accessed 15 May 2020).

control, since obscurity, biases or profiling are much more difficult to detect, at least in some of the stages of the life cycle of the AI development.

In this part of The Guidelines, we try to provide support for both supervised and unsupervised AI. We are aware that it is almost impossible to provide advice on every possible situation. However, we hope we will be able to highlight the fundamentals and include useful additional sources of information. Finally, we fully understand that some experts could consider that some of the recommendations we make could be moved from one step to another. Furthermore, some of them could apply to several different steps. Therefore, we strongly recommend that they adapt these Guidelines to their best convenience and knowledge.

The structure of the document is easy to follow. First, we introduce a quote to the chapter by Colin Shearer,[124] followed by a description of the tasks involved in each concrete stage of the process, according to the same author. Next, we introduce some recommendations that should be implemented at that point. References to other chapters of The Guidelines are highlighted, while references to other parts of this chapter are cross-referenced. Finally, the annexes include references to some tools that might serve the purposes of this part of The Guidelines. Annex I shows the recommendations for auditing AI tools elaborated by the Spanish Data Protection Agency. Annex II is more specific, as it refers to the use of AI in the healthcare sector. However, it is an excellent guide for those who are willing to develop an AI tool in that sector. In the future, we will try to incorporate more annexes, as soon as an efficient mechanism for doing so is produced.

1 Business understanding

"The initial business understanding phase focuses on understanding the project objectives from a business perspective, converting this knowledge into a data mining problem definition, and then developing a preliminary plan designed to achieve the objectives. In order to understand which data should later be analyzed, and how, it is vital for data mining practitioners to fully understand the business for which they are finding a solution. The business understanding phase involves several key steps, including determining business objectives, assessing the situation, determining the data mining goals, and producing the project plan."[125]

[124] Shearer, C. (2000) 'The CRISP-DM model: the new blueprint for data mining', *Journal of Data Warehousing* 5(4): 13-23. Available at: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf (accessed 15 May 2020).

[125] Shearer, C. (2000) 'The CRISP-DM model: the new blueprint for data mining', *Journal of Data Warehousing* 5(4): 13-23, p.14.Available at: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf (accessed 15 May 2020).

1.1 Description

This general objective involves four main tasks:

1. Determine the business objectives:
    a. Uncover the primary business objective as well as the related questions the business would like to address.
    b. Determine the measure of success.
2. Assess the situation:
    a. Identify the resources available to the project, both, material and personal.
    b. Identify what data is available to meet the primary business goal.
    c. List the assumptions made in the project.
    d. List the project risks, list potential solutions to those risks, create a glossary of business and data mining terms, and construct a cost-benefit analysis for the project.
3. Determine the data mining goals:
    a. Decide what level of predictive accuracy is expected to consider the project successful.
4. Produce a project plan:
    a. Describe the intended plan for achieving the data mining goals, including outlining specific steps and a proposed timeline, an assessment of potential risks, and an initial assessment of the tools and techniques needed to support the project.

1.2 Main actions that need to be addressed

1.2.1 Deciding about your business objectives

AI developers should know from the beginning what they expect the tool to be capable of doing. The more inaccurate they are about these expectations, the more difficult it becomes to determine the precise purposes of the processing (see "Prerequisites to lawfulness specified, explicit purposes" in the 'Lawfulness, fairness and transparency' subsection of the "Principles" in Part II of these Guidelines). If we keep in mind that controllers must make the purposes of processing explicit, that is, "revealed, explained or expressed in some intelligible way",[1] accurate expectations are strongly recommended. However, one must distinguish between the different stages of the life cycle of an AI development. In the training stage, the use of large amounts of data might be essential to estimate the concrete utility of the tool. Therefore, processing big datasets might be acceptable even though the specific end (developing the AI tool) is not so precise. This, of course, would not be so easily acceptable if we were in the last stage of the process, that is, the deployment and use of the tool. If, at that moment, the

controller would need to use a large amount of data, a much more detailed justification would be needed.

In any case, it is necessary to highlight that some key ideas must be kept in mind from the very beginning. For instance, deciding the expected level of predictive accuracy, in order to consider the project a success, is essential to assess the amount of data that will be needed to develop the AI tool or the nature of that data. The level of predictability or precision of the algorithm, the validation criteria to test it, the maximum quantity or the minimum quality of the data that will be necessary to use it in the real world, etc., are fundamental features of an AI development. These key decisions should be considered from the first stage of the solution's life cycle. This will be extremely helpful to implement a data protection by design policy (see the see "Data protection by design and by default" within Part II section "Main concepts" of these Guidelines).

Thus, the AI developer should fix acceptable thresholds of false positives/negatives or ranges, depending on the use case and then perform a utility balance. The AI developer must be aware that determining the expected level of accuracy is clearly linked to the amount of data needed. It is not the same to develop, for example, a product for healthcare or for recommending TV series. In addition, even within the health sector, it is not the same to develop a tool capable of performing a first triage (that is, recommending whether a primary care physician or a specialist should intervene) or a solution that aims to support radiologists in diagnosis. Depending on what the mechanism is intended to do, higher or lower accuracy requirements will be adopted.

If an acceptable level of accuracy could be reached by using considerably less personal data than required by a higher level of accuracy, then this should be strongly considered. Furthermore, AI developers must keep in mind that any marginal increase in terms of accuracy of the prediction sometimes calls for a significant increase in the amount of personal data needed.[2] Therefore, if they are considering a fundamental modification in the level of accuracy of the prediction required, they should carefully consider if this works well with the data minimization principle (see "Data minimization principle" within Part II section "Principles" of these Guidelines)..

1.2.2 Opting for the technical solution

In general, AI developers should always provide for the development of more understandable algorithms over less understandable ones (see GDPR provisions: Transparency" section). Trade-offs between the explainability/transparency and best performance of the system must be appropriately balanced based on the context of use. For instance, in healthcare the accuracy and performance of the system may be more important than its explainability, whereas, in policing, explainability is much more crucial to justify behaviors and outcomes of law enforcement. In other areas, such as recruitment, both accuracy and explainability are similarly valued.[126] If a service can be offered through both, by an easy to understand and an opaque algorithm, that is, when there is no trade-off between explainability and performance, the controller should opt

[126] SHERPA project (2019) Guidelines for the ethical development of AI and big data systems: an ethics by design approach. SHERPA, p.26. Available at: www.project-sherpa.eu/wp-content/uploads/2019/12/development-final.pdf (accessed 15 May 2020).

for the one that is more interpretable (se "Lawfulness, fairness and transparency" section in the "Principles" in Part II of these Guidelines).

Box 13: Interpreting interpretability

Even though interpretability seems to be recommended, it must be acknowledged that this is not a clear concept. The academic literature shows different motivations for interpretability and, more importantly, offers myriad notions of what attributes render models interpretable. It is still unclear what "interpretation" brings together. At first sight it seems reasonable to suppose that simple, lineal algorithms are easier to understand. However, "for some kinds of post-hoc interpretation, deep neural networks exhibit a clear advantage. They learn rich representations that can be visualized, verbalized, or used for clustering. Considering the desiderata for interpretability, linear models appear to have a better track record for studying the natural world but we do not know of a theoretical reason why this must be so. Conceivably, post-hoc interpretations could prove useful in similar scenarios." Therefore, it is hard to arrive at specific recommendations on which type of models should be preferred on the basis of their "interpretability"[127]

1.2.3 Implementing a training program

This action is one of the most important pieces of advice to be considered from the very first moment of an AI business development. Algorithm designers (developers, programers, coders, data scientists, engineers), who occupy the first link in the algorithmic chain, are likely to be unaware of the ethical and legal implications of their actions. Furthermore, one of the main problems that AI raises is that it generally uses personal data that are included in large datasets. This somehow blurs the relationship between the data and the data subject, leading to violations of the regulations that rarely occur when the controller and the subject have a direct relationship.[128] This could bring consequences in terms of adequate compliance with data protection standards. It is paramount that these key workers have the fullest possible awareness of the ethical and social implications of their work, and of the very fact that these can even extend to societal choices,[129] even though the 'rogue engineering' alibi can hardly function after the Google Street View case.[130]

In order to avoid that the misrepresentation of the ethical and legal issues provokes unwanted consequences, there are two main courses of action that can be adopted. First,

---

[127] Lipton, Z.C. (2017) 'The mythos of model interpretability', 2016 ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. Available at: https://arxiv.org/pdf/1606.03490.pdf (accessed 15 May 2020).

[128] Kuyumdzhieva, A. (2018) 'Ethical challenges in the digital era: focus on medical research', pp. 45-62 in: Koporc, Z. (ed.) *Ethics and integrity in health and life sciences research*. Emerald, Bingley.

[129] CNIL (2017) How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence. Commission Nationale de l'Informatique et des Libertés, Paris, p.55. Available at: www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf (accessed 15 May 2020).

[130] See, for instance: https://www.slashgear.com/googles-rogue-engineer-street-view-excuse-blown-apart-30225200/

developers might try to ensure that algorithm designers are able to understand the implications of their actions, both for individuals and society, and be aware of their responsibilities by learning to show continued attention and vigilance.[131] In that sense, an optimal training for all subjects involved in the project (developers, programers, coders, data scientists, engineers, researchers) even before it starts could be one of the most efficient tools to save time and resources in term of compliance with data protection regulation. Thus, implementing basic training programs that include at least the fundamentals of the Charter of Fundamental Rights, the principles exposed in Article 5 of the GDPR, the need for a legal basis for processing (including contracts between the parties), etc.

However, training people who have never been in touch with data protection issues might be hard. An alternative policy is the involvement of an expert on data protection, ethical and legal issues in the development team, so as to create an interdisciplinary team. This might be done by hiring an expert for this purpose (an internal worker or an external consultant) to design the strategy and the subsequent decisions on personal data required by the development of the tools, with the close involvement of the Data Protection Officer.

Adopting adequate measures in terms of ensuring confidentiality, integrity and availability of data is also strongly recommendable (see "Measures in support of confidentiality" in the "Integrity and confidentiality" subsection of the "Principles" in Part II of these Guidelines).

1.2.4 **Designing legitimate data processing tools** According to Article 5(1)(a) of the GDPR, personal data shall be "collected for specific, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes". The concept of legitimacy is not well defined in the GDPR, but the Article 29 Working Party stated that legitimacy involves that data must be processes "in accordance with the law", and "law" should be understood as a broad concept that includes "all forms of written and common law, primary and secondary legislation, municipal decrees, judicial precedents, constitutional principles, fundamental rights, other legal principles, as well as jurisprudence, as such 'law' would be interpreted and taken into account by competent court".[132]

Therefore, it is a wider concept than lawfulness. It involves compliance with the main values of the applicable regulation and the main ethical principles at stake. For instance, some concrete AI developments will need the intervention of an ethics committee. In other cases, guidelines or any other kind of soft regulation might be applicable. AI developers should ensure adequate compliance with this requirement by designing a plan form this preliminary stage of the lifecycle of the tool (see "Legitimacy and lawfulness" in the "Lawfulness, fairness and transparency" subsection of the 'Principles' in Part II of these Guidelines).

---

[131] Ibid., p.55.

[132] Article 29 Working Party (2013) Opinion 03/2013 on purpose limitation Adopted on 2 April 2013, WP203. European Commission, Brussels, p.20. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf (accessed: 15 May 2020)

1.2.5 Adopting a risk-based thinking approach

Controllers should minimize the risks to data subjects' rights, interests, and freedoms. To this purpose, they should work on a risk-based approach (see "Integrity and confidentiality principle" in Part II, section "Principles").

The risk-based approach of data protection law requires controllers to comply with their obligations and implement appropriate measures in the context of their particular circumstances – the nature, scope, context and purposes of the processing they intend to do, and the risks this poses to individuals' rights and freedoms. Their compliance considerations therefore involve assessing the risks to the rights and freedoms of individuals and taking judgements as to what is appropriate in those circumstances. In all cases, controllers need to ensure that they comply with data protection requirements (see "Accountability principle" in Part II, section "Principles").

Risk-based thinking with regard to confidentiality of data, or a risk-based approach to questions of what harm may be done to people, must be included from the first steps of the process. It might be too late if it is only considered later. To manage the risks to individuals that arise from the processing of personal data in AI tools, it is important that controllers develop a mature understanding and articulation of fundamental rights, risks, and how to balance these and other interests. Ultimately, it is necessary for controllers to assess the risks to individuals' rights that the use of AI poses, and determine how they need to address these and establish the impact this has on their use of AI.[133] To this purpose, there are two key factors that must be considered:[134]

- Risks arising from the processing itself, such as the emergence of biases associated with profiling or automated decision-making systems (see **GDPR provisions**).
- Risks arising from the processing in relation to the social context and the side effects indirectly related to the object of processing that may occur.

[133] ICO (2020) AI auditing framework - draft guidance for consultation. Information Commissioner's Office, Wilmslow, p.13-14. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (accessed 15 May 2020).

[134] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.30. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 15 May 2020).

Box 14: The importance of software providers in terms of security

In August 2015, an Indiana medical software company reported to the federal government that its networks had been hacked and the private information of 3.9 million people exposed. That included personal data such as names, addresses, birthdates, Social Security numbers and health records. According to IBM X-Force Research, this was one of the biggest healthcare data breaches in recent years. According to the company, the attack was detected nineteen days after the authors gained unauthorized access to its network. Clients were only notified almost a month after the attack began.[135]

In addition, controllers must ensure that appropriate technical and organizational measures are implemented to eliminate, or at least mitigate the security risk, reducing the probability that the identified threats will materialize, or reducing their impact. The general description of the technical and organizational security measures must become a part of the records of processing, where possible (Article 30(1) (g) for controllers, and 30(2)(d) for processors) and all implemented measures are part of the DPIA, as supporting remediation measures to limit risk. Finally, once the selected measures are implemented, the remaining residual risk should be assessed and kept under control. Both the risk analysis and the DPIA are the tools that apply.

A DPIA is very often compulsory in the case of AI development (see "In what cases must I carry out a DPIA" in subsection "Data Protection Impact Assessment" of the "Main Tools and Actions", Part II). It depends on whether the risks associated with the processing are high or not, according to Article 35(3) of the GDPR. However, it is highly recommended as it supports accountability. In case of doubt, consultation of the competent supervisory authority prior to processing is highly recommended. Finally, do not forget that when using big data and AI it is hard to foresee what the future risks will be, so doing assessment of ethical implications will not be sufficient to address all possible risks. Therefore, it is important to consider having a reassessment of risks and also highly recommendable to integrate a more dynamic way of assessing research risks. Do not hesitate to perform additional DPIAs in other stages of the process if need be.

1.2.6 Preparing the documenting of processing

Whoever processes personal data (including both controllers[136] and processors[137]) needs to document their activities primarily for the use of qualified/relevant Supervisory Authorities.[138] This must be done through records of processing that are maintained

[135] IBM X-Force® Research (2017) Security trends in the healthcare industry: Data theft and ransomware plague healthcare organizations. IBM Security, Somers, NY, p.7. Available at: www.ibm.com/downloads/cas/PLWZ76MM (accessed: 17 May 2020).

[136] See Article 30(1) of the GDPR.

[137] See Article 30(2) of the GDPR.

[138] See Articles 58(1)(a), 30(4) and 5(2) of the GDPR.

centrally by the organization across all its processing activities, and additional documentation that pertains to an individual data processing activity (see "Documentation of processing" in the "Main Tools and Actions" section of Part II of these Guidelines). This preliminary stage is the perfect moment to set up a systematic way of collecting the necessary documentation, since it will be the time when the organization conceives and plans the processing activity[139].

Indeed, controllers should create a Data Protection Policy that allows the traceability of information (if approved codes of conduct exist, these should be implemented; see the "Economy of scale for compliance and its demonstration" subsection in the "Accountability" section of the "Principles" in Part II of these Guidelines). This policy should also make the responsibilities assigned to processors clear, and include in the processing agreement tasks that will be delegated to it in relation to the execution of data subjects' rights. AI developers should always remember that Article 32(4) of the GDPR clarifies that an important element of security is to ensure that employees act only on instruction and as instructed by the controller (see "Integrity and confidentiality" in Part II, section "Principles").

Controllers must always keep in mind that the development of AI tools often involves the use of different datasets. The traceability of the processing, the information about possible re-use of data, and the use of data pertaining to different datasets in different or in the same stages of the life cycle must be ensured by the records.

1.2.7 Documenting of processing

As stated in the requirements and acceptance tests for the purchase and/or development of the employed software, hardware, and infrastructure (see the subsection of the "Documentation of processing" section), the risk evaluation and the decisions taken "have to be documented in order to comply with the requirement of data protection by design" (of Article 25 of the GDPR).

Finally, the controllers should always be aware that, according to Article 32(1)(d) of the GDPR, data protection as a process. Therefore, **they should test, assess, and evaluate the effectiveness of technical and organizational measures regularly.** Procedures that serve controllers to identify changes that would trigger the revisit of the DPIA should be created at this moment. Whenever possible, controllers should try to impose a dynamic model of monitoring the measures at stake (see "Integrity and confidentiality" in Part II of these Guidelines, section "Principles").

Box 15: The extreme difficulty of accountability in AI development

Even though accountability is a necessary goal and assigning responsibilities to a specific processor is absolutely necessary, controllers must always be conscious that AI functioning can make it extremely difficult to monitor a system. As the CNIL stated, "the question of where accountability and decision-making can be set up is to be approached in a slightly different way when dealing with machine learning systems". Therefore, controllers should better think more in terms of a chain of accountability,

[139] Article 25(1) of the GDPR calls this "the time of the determination of the means for processing".

from the system designer right through to its user, via the person who will be feeding the training data into this system. The latter will operate differently depending on such input data.

On this subject, one could mention Microsoft's chatbot Tay. It was shut down a mere twenty-four hours after its release when, learning from the posts of social media users, it had begun to tweet racist and sexist comments of its own. Needless to say, working out the precise share of responsibility between these different links in the chain might be a laborious task.[140]

1.2.8 Checking regulatory framework

The GDPR includes a specific regulatory framework regarding processing for the purposes of scientific research (see "Data protection and scientific research" in Part II, section "Main Concepts").[141] If the AI development could be considered as scientific research, the "Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes" (Article 89(2)). Furthermore, according to article 5 (b) "further processing of the data gathered, in accordance with Article 89(1), would not be considered to be incompatible with the initial purposes ('purpose limitation')". There are some other particular exceptions to the general framework applicable to processing for research purposes (such as storage limitation) that should also be taken into account. Nevertheless, AI developers should be aware of the concrete regulatory framework that applies to their research. It might include important changes depending on their national regulations. Consultation with their DPOs is highly recommended for this purpose.

1.2.9 Defining data storage policies

According to Article 5(1)(e) of the GDPR, personal data should be "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed". This requisite is twofold. On one hand, it relates to identification: data should be stored in a form which permits identification of data subjects for no longer than necessary. Consequently, AI developers should implement policies devoted to avoiding identification as soon as it is not necessary for processing. This involves the adoption of adequate measures to ensure that at any moment, only **the minimal degree of identification that is necessary to**

[140] CNIL (2017) How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence. Commission Nationale de l'Informatique et des Libertés, Paris, p.29. Available at: www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf (accessed 15 May 2020).

[141] This specific framework also includes historical research purposes or statistical purposes. However, ICT research is not usually related to these purposes. Therefore, we will not analyze them.

fulfil the purposes must be used (see "Storage limitation" in Part II, section "Principles").

On the other hand, data storage implies that data can only be stored for a **limited period**: the time that is strictly necessary for the purposes for which the data are processed. However, the GDPR permits 'storage for longer periods' if the sole purpose is scientific research (or archiving in the public interest, historical research or statistical purposes) (see "Data protection and scientific research" in Part II, section "Main Concepts").

In the case of AI development, this exception raises the risk that developers decide to keep the data longer than strictly needed, so as to ensure that they are available for reasons other than the original purposes they were collected for. The controllers should be aware that even though the GDRP might allow storage for longer periods, **they should have a good reason to opt for such an extended period (**see "Temporal aspect" in the 'Storage limitation' subsection of the "Principles" in Part II). Provided that a real risk comes from the lack of respect of the purpose limitation principle, **the compatibility test should be part of any potential reuse of the data**.

The intention of the lawmaker appears to have been to dissuade unlimited storage even in this special regime, and guards against scientific research as a pretext for prolonged storage for other, private, purposes. If in doubt, the controller should consider whether a new legal basis is appropriate. Therefore, storage periods should be proportionate to the aims of the processing: "In order to define storage periods (timelines), criteria such as the length and the purpose of the research should be taken into account. It has to be noted that national provisions may stipulate rules concerning the storage period as well."[142]

Thus, if controllers do not need the data, and there are no compulsory legal reasons that oblige them to preserve the data, they should better anonymize or delete them. Researchers should consult their DPOs if they are willing to store data for a long-lasting period and be aware of the applicable national regulation. This could also be an excellent moment to **envisage time limits for erasure of the different categories of data and document these decisions** (see the "Accountability" section of the "Principles" in Part II of these Guidelines.

1.2.10 Appointing a Data Protection Officer

The appointment of a DPO is one of the best steps that can be taken by the controller to properly implement measures that ensure compliance with the rights of the data subjects. Appointing a DPO is not a necessary consequence of operating with AI tools. It is undeniable, however, that appointing a DPO is only compulsory if conditions settled by Article 37(1) (b) or (c) apply. Therefore, it is not the case that all AI

[142] EDPS (2020) Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak Adopted on 21 April 2020. European Data Protection Supervisor, Brussels, p.10. Available at https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf (accessed 23 April 2020).

developers should appoint a DPO. However, **it is always recommendable to proceed to do so, at least in terms of transparency** (see the "Transparency" section of the "Principles" in Part II)**.**

In any case, the data responsible should elaborate this by outlining the role of the DPO in relation to the overall management of the project, ensuring that the role of the DPO is not marginal, but cemented into the decision-making processes of the organization/project. They should make clear what that role could be in terms of oversight, decision-making and similar.

2 Data understanding

"The data understanding phase starts with an initial data collection. The analyst then proceeds to increase familiarity with the data, to identify data quality problems, to discover initial insights into the data, or to detect interesting subsets to form hypotheses about hidden information. The data understanding phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality".[143]

2.1 Description

At this stage, initial data collection takes place and an initial study of the data is performed. It involves four sequential tasks:

- Collect initial data
- Describe data
- Analyze data
- Verify data quality.

All of these tasks are aimed at identifying the available data. At this stage, developers need to be aware of the data they will have to work with and start making decisions on the way in which main principles related to data protection will be implemented.

2.2 Main actions that need to be addressed

At this stage, there are a huge number of fundamental issues related to the protection of personal data that need to be addressed. Depending on the decisions made, principles

[143] Shearer, C. (2000) 'The CRISP-DM model: the new blueprint for data mining', *Journal of Data Warehousing* 5(4): 13-23, p.15. Available at: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf (accessed 15 May 2020).

such as data minimization, privacy by design or by default, lawfulness, fairness and transparency, etc. will be adequately settled.

2.2.1 Type of data collected

According to the GDPR, the controller "shall implement appropriate technical and organizational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons"[144] (see the "Data protection by design and by default", section in the "Main Concepts" of Part II). This must be kept in mind particularly during this stage, since decisions about the type of data that will be used are often taken at this moment.

Controllers must consider that it is always better to avoid using personal data if this is possible. Indeed, according to the data minimization principle, the use of personal data should be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed. Therefore**, if the same purpose could be reached without using personal data, processing should be avoided.**

In a second level of precaution, **if developers need to use personal data, they should try to avoid using special category data.** Sometimes this is doable, sometimes it is not. This often depends on the area of application of the model. It is not the same working on a model that will be used for the analysis of the influence of epigenetics in human health, a model used to monitor an epidemic outbreak or a model that will serve to target advertisements accurately. If such special category data are finally used, controllers must consider the regulations applying to their processing and the necessary application of appropriate safeguards, able to protect the data subjects' rights, interests and freedoms. Proportionality between the aim of research and the use of special categories of data must be guaranteed. Furthermore, controllers must ensure that their Member States regulation do not protect genetic, biometric and health data by introducing further conditions or limitations, since they are empowered to do so by the GDPR.

If personal data are necessary, the AI developer should, at least, try to reduce the amount of data considered as much as possible (see the "Data minimization" section in the "Principles" in Part II). They should always remember that they can only process data if the processing is adequate and relevant. Therefore, they should avoid using excessive amount of personal data. Too often, this is easier to do that it seems. As the Norwegian Data Protection Agency states, "[i]t is worth noting that the quality of the training data, as well as the features used, can in many instances be substantially more important than the quantity. When training a model, it is important that the selection of training data is representative of the task to be solved later. Huge volumes of data are of little help if they only cover a fraction of what the model will subsequently be working

[144] Article 24 of the GDPR.

on."[145] Therefore, it is particularly important **not to collect unnecessary data**. Correct labelling could be a nice antidote against unnecessary collection. Note that **if data is already stored, selection involves deleting unnecessary data elements.**

The developer should always try to avoid the 'Curse of Dimensionality'; that is, "a poor performance of algorithms and their high complexity associated with data frame having a big number of dimensions/features, which frequently make the target function quite complex and may lead to model overfitting as long as often the dataset lies on the lower dimensionality manifold."[146] To this purpose**, having an expert able to select relevant features might be extremely important.** This would contribute to significantly reduce the amount of personal data used without losing quality. This should not be difficult if the data scientist is well acquainted with the dataset and the meanings of its numerical features. Under such conditions, it would be easy to determine if some of the variables are needed or not. However, it is possible to perform such an approach only in the case where the dataset is easily interpreted and the dependencies between the variables are well known. Therefore, the developer will need a smaller amount of data if they have been adequately classified. Smart data might be much more useful than big data. Of course, this might involve a huge effort in terms of unification, homogenization, etc., but it will help to implement the principle of data minimization (see "Data minimization principle" within Part II section "Principles" of these Guidelines) in a much more efficient way.

Furthermore, the controllers should try to **limit the resolution of the data** to what is minimally necessary for the purposes pursued by the processing. They should also **determine an optimal level of data aggregation** before starting the processing (see the "Adequate, relevant and limited" section of the 'Data minimization" section of the "Principles" in Part II).

Data minimization might be complex in the case of deep learning, where discrimination by features might be impossible. There is an efficient way to regulate the amount of data gathered and increase it only if it seems necessary: the learning curve[147]. The developer should start by gathering and using a restricted amount of training data, and then monitor the model's accuracy as it is fed with new data.

Box 16: A data minimization practice that was not adequately implemented

A tool developed by the Norwegian Tax Administration to filter tax returns for errors tested 500 variables in the training phase. However, only 30 were included in the final AI model, as they proved most relevant to the task at hand. This means that they could

[145] Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf (accessed 15 May 2020).

[146] Oliinyk, H. (2018) Why and how to get rid of the curse of dimensionality right (with breast cancer dataset visualization). Towards Data Science, 20 March. Available at: https://towardsdatascience.com/why-and-how-to-get-rid-of-the-curse-of-dimensionality-right-with-breast-cancer-dataset-7d528fb5f6c0 (accessed 15 May 2020).

[147] Ng, R. (no date) Learning curve. Available at: www.ritchieng.com/machinelearning-learning-curve/ (accessed 15 May 2020).

have probably avoided collecting so much personal data if they had performed a better selection of the variables that were relevant from the very beginning.[148]

2.2.2 Selecting appropriate legal basis for processing

Controllers should decide the legal basis that they will use for processing before starting it, document their decision privacy notice (along with the purposes) and include the reasons why they have made such choices (see the "Accountability" section in the "Principles" in Part II). In principle, they should select the **legal basis that most closely reflects the true nature of their relationship with the individual and the purpose of the processing.** This decision is key, since changing the legal basis for processing is not possible if there are not solid reasons that justify it (see the "Purpose limitation" section of the "Principles" in Part II).

In principle, consent is one of the most common legal grounds for processing (see the "Consent" section of the "Main Concepts" in Part II). However, it involves certain risks. Namely, consent is always linked to specific purposes. Therefore, 'widening' the purposes of processing beyond data subjects' explicit consent shall be rendered unlawful processing. In order to determinate whether further processing is compatible or not with the original processing controllers should make use of the criteria included in Article 6(4) of the GDPR (see the "When are purposes compatible?" subsection in the "Purpose limitation" section). As mentioned, processing for scientific or historical research purposes or statistical purposes shall not be considered incompatible with the initial purposes (see the "Data protection and scientific research" section in the "Main Concepts" in Part II).

The most common alternative grounds for processing data in AI are legitimate interests, performance of contract and legal obligation or vital interest. All of them involve specific characteristics that must be carefully analyzed.

2.2.3 Checking legitimate dataset usage

Datasets can be obtained in different ways. Firstly, the developer might opt for acquiring or gaining access to a database that has already been built by someone else. If this is the case, the controller should be particularly careful, since there are a lot of legal issues that relate to the acquisition of access to database (see the "Purchasing access to a database" section in the "Main Tools and Actions" of Part II).[149]

Secondly, the most common alternative to this consists of building a database. Quite obviously, in this case controllers have to ensure that they comply with all legal

[148] Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf (accessed 15 May 2020).

[149] Yeong Z.K. (2019) Legal issues in AI deployment. Law Gazette, February. Available at: https://lawgazette.com.sg/feature/legal-issues-in-ai-deployment/ (accessed 15 May 2020).

requirements imposed by the GDPR to create a database (see the "Creating a database" section within the "Main Tools and Actions" in Part II of these Guidelines).

Thirdly, sometimes developers choose an alternative path. They **mix licensed data from third parties with each other or with the controllers' own dataset so as to create a huge training dataset and another one for validation purposes.** This could bring some issues, such as for example the possibility that the combination of these personal data provides some additional information about the data subjects. For instance, it could allow the controller to identify data subjects, something that was previously not possible. That could involve de-anonymizing anonymized data and creating new personal information that was not contained in the original data set, a circumstance that would bring dramatic ethical and legal issues. Therefore, re-identification must be tested through methods such as k-anonymity, l-diversity or t-closeness techniques.[150]

Another common issue is that the original basis for processing the data gathered in each dataset is diverse. If a controller merges the datasets and then one of the legal bases is no longer applicable, that controller faces a terrible situation. For instance, if one of the databases was built on the basis of consent and some of the data subjects withdraw their consent, the controller will have to delete them from the merged dataset. This might be really hard to do in practice.

Furthermore, new information produced in this way may also be based on probabilities or conjectures, and therefore be false, or contain biases in the portrayal of persons (see the "**GDPR provisions**" section).[151] Therefore, controllers should try to avoid such consequences by ensuring that merging datasets do not work against data subjects' rights and interests.

Finally, if controllers use several datasets that pursue different purposes, they should implement adequate measures to separate the different processing activities. Otherwise they could easily use data collected for on purpose to different activities. This might bring issues related to the purpose limitation principle.

3 Data preparation

"The data preparation phase covers all activities to construct the final data set or the data that will be fed into the modeling tool(s) from the initial raw data. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for

[150] Rajendran, K., Jayabalan, M. and Rana, M.E. (2017) 'A study on k-anonymity, l-diversity, and t-closeness techniques focusing medical data', International Journal of Computer Science and Network Security 17(12): 172-177.

[151] SHERPA project (2019) Guidelines for the ethical development of AI and big data systems: an ethics by design approach. SHERPA, p.38. Available at: www.project-sherpa.eu/wp-content/uploads/2019/12/development-final.pdf (accessed 15 May 2020).

modeling tools. The five steps in data preparation are the selection of data, the cleansing of data, the construction of data, the integration of data, and the formatting of data."[152]

3.1 Description

This stage includes all activities needed to construct the final dataset that is fed into the model, from initial raw data. It involves the following five tasks, not necessarily performed sequentially.

1. Select data. Decide on the data to be used for analysis, based on relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types.
2. Clean data. Raise data quality to a required level, for example by selecting clean subsets of the data, insertion of defaults, and estimation of missing data by modeling.
3. Construct data. The construction of new data through the production of derived attributes, new records, or transformed values for existing attributes.
4. Integrate data. Combine data from multiple tables or records to create new records or values.
5. Format data. Make syntactic modifications to data that might be required by the modeling tool.

3.2 Main actions that need to be addressed

3.2.1 Ensuring accuracy of personal data

According to the GDPR, data must be accurate (see the "Accuracy" section in the "Principles" in Part II). This means that data are correct and up to date, but also refers to the accuracy of the analytics performed. The EDPB has highlighted the importance of the accuracy of the profiling or the (not exclusively) automated decision-making process at all stages (from the collection of the data to the application of the profile to the individual).[153]

---

[152] Shearer, C. (2000) 'The CRISP-DM model: the new blueprint for data mining', *Journal of Data Warehousing* 5(4): 13-23, p.16. Available at: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf (accessed 15 May 2020).

[153] *Guidelines* on *Automated* individual *decision-making and Profiling* for the purposes of Regulation 2016/679 (wp251rev.01). 22/08/*2018, p.* 13; Ducato, Rossana, Private Ordering of Online Platforms in Smart Urban Mobility The Case of Uber's Rating System, CRIDES Working Paper Series no. 3/20202 February 2020 Updated on 26 July 2020, p. 20-21, at: https://poseidon01.ssrn.com/delivery.php?ID=247104118003073117118086021112071111102048023015008020118084071112086000027097102088036101006014057116105116119119026079007006118

Controllers are responsible of ensuring accuracy of personal data. Therefore, once they have finished with the collection of personal data, they should implement adequate tools to guarantee the accuracy of those data. This basically involves the implementation of technical and organizational measures that will ensure that this principle is applicable (see the "Related technical and organizational measures" subsection in the "Accuracy" section of the "Principles" in Part II). If personal data proceed from data subjects, the controller can assume that they are accurate (unless the person responsible considers that the data subject might have a reason to provide inaccurate data). If personal data have not been collected from the data subject, controllers are obliged "to verify the accuracy of the obtained data, at least in respect of fitness for the declared purposes of processing and to any negative consequences that inaccuracies may have for data subjects." (see the "How is inaccuracy of data discovered?" subsection in the "Accuracy" section of the "Principles" in Part II). In any case, accuracy requires an adequate implementation of measures devoted to facilitate the data subjects' right to rectification (see "Right to rectification" in Part II, section "Data Subjects' Rights").

3.2.2 Focussing on profiling issues

In the case of a database that will serve to train or validate an AI tool, there is a particularly relevant obligation to inform the data subjects that **their data might cause automated decision-making or profiling on them, unless controllers can guarantee that the tool will in no way produce these consequences.**

Even though automatic decision-making can hardly happen in the context of research, developers should pay attention to this issue.

Profiling, on the other hand, might bring some problems to AI development.

This is due to a simple reason: the process of profiling is "often invisible to the data subject. It works by creating derived or inferred data about individuals – 'new' personal data that has not been provided directly by the data subjects themselves. Individuals have differing levels of comprehension and may find it challenging to understand the complex techniques involved in profiling and automated decision-making processes."[154] Thus, "if the controller envisages a 'model' where it takes solely automated decisions having a high impact on individuals based on profiles made about them and it cannot rely on the individual's consent, on a contract with the individual or on a law authorizing this, the controller should not proceed."[155] Risk for the individual's rights, interests and freedoms is a very important factor that must always be considered. It is not the same sort of profiling to make a decision on someone's taste for TV series, compared to profiling devoted to approve their health insurance policy. Thus, if

---

0440330550001140231060070761150960730240070940810020780640980280910930030780950990821081130860981200010790151230270831250 24&EXT=pdf&INDEX=TRUE

[154] Article 29 Working Party (2017) Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679. Adopted on 3 October 2017 as last Revised and Adopted on 6 February 2018. European Commission, Brussels, p.9. Available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053 (accessed 15 May 2020)

[155] Ibid., p.30.

processing presents risks to individuals' fundamental rights and freedoms, the controllers must ensure that they can address these risks and meet the requirements.

If processing and/or automated decision-making happens, data subjects **must be informed adequately about that processing and the way the algorithm works.** In other words, their right to information must be satisfied in application of the lawfulness, fairness and transparency principle. This means that, at least, controllers have to tell the data subject that "they are engaging in this type of activity, provide meaningful information about the logic involved and the significance and envisaged consequences of the profiling for the data subject."[156]

The information about the logic of a system and explanations of decisions should give individuals the necessary context to decide whether, and on what grounds, they would like to request human intervention. In some cases, insufficient explanations may prompt individuals to resort to other rights unnecessarily, or to withdraw their consent. Requests for intervention, expression of views, or contests are more likely to happen if individuals do not feel they have a sufficient understanding of how the decision was reached.[157]

Finally, a controller must always remember that according to Article 22(3), automated decisions that involve special categories of personal data are permitted only if the data subject has consented, or if they are conducted on a legal basis (see the '**Human agency and oversight**' section in this part of the Guidelines). This exception applies not only when the observed data fit into this category, but **also if the alignment of different types of personal data can reveal sensitive information about individuals or if inferred data enter into that category.** In all of these cases, we must talk about processing of special categories of personal data. Indeed, a study capable of inferring special categories of data is subject to the same legal obligations pursuant to the GDPR as if sensitive personal data had been processed from the outset. If profiling infers personal data that were not provided by the data subject, the controllers should ensure that the processing is not incompatible with the original purpose, they have identified a lawful basis for the processing of the special category data, and they inform the data subject about the processing.[158]

Box 17: Example of inferring special categories data

Research showed that in 2011, easily accessible digital records of behavior, Facebook 'likes', could be used to automatically and accurately predict a range of highly sensitive

[156] Ibid., pp.13-14.

[157] ICO (2020) AI auditing framework - draft guidance for consultation. Information Commissioner's Office, Wilmslow, p.94. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (accessed 15 May 2020).

[158] Article 29 Working Party (2018) Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. Adopted on 3 October 2017 as last Revised and Adopted on 6 February 2018. European Commission, Brussels, p.15. Available at: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053 (accessed 15 May 2020).

personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis was based on a dataset of over 58,000 volunteers who provided their Facebook 'likes', detailed demographic profiles, and the results of several psychometric tests. The model 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. For the personality trait 'Openness', prediction accuracy was close to the test–retest accuracy of a standard personality test. The authors provided examples of associations between attributes and Likes and discuss implications for online personalization and privacy.[159]

Performing a DPIA is compulsory if there is a real risk of unauthorized profiling or automated decision-making. Article 35(3) (a) of the GDPR states the need for the controller to carry out a DPIA in the case of a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person. Controllers should be aware that, at the present moment, each country has submitted their lists of when a DPIA is required to the EDPB. If the controller is within the EEA, this list should also be locally verified[160] **(**see "DPIA" within Part II section "Main actions and tools" of these Guidelines).

According to Article 37(1)(b) and (5) of the GDPR, controllers shall designate a data protection officer where "the core activities of the controller or the processor consist of processing operations which, by virtue of their nature, their scope and/or their purposes, require regular and systematic monitoring of data subjects on a large scale." Controllers are also required to keep a record of all decisions made by an AI tool as part of their accountability and documentation obligations. This should also include whether an individual requested human intervention, expressed any views, contested the decision, and whether a decision has been altered as a result[161] (see the "Accountability principle" section in the "Principles" within Part II).

Some additional actions that might be extremely useful to avoid automated decision-making are as follows:[162]

---

[159] Kosinski, M., Stillwell, D. and Graepel, T. (2013) 'Digital records of behavior expose personal traits', *Proceedings of the National Academy of Sciences* 110 (15): 5802-5805, DOI: 10.1073/pnas.1218772110.

[160] EDPB (2019) Data Protection Impact Assessment. European Data Protection Board, Brussels. Available at: https://edpb.europa.eu/our-work-tools/our-documents/topic/data-protection-impact-assessment-dpia_es (accessed 3 June 2020).

[161] ICO (2020) Guidance on the AI auditing framework - draft guidance for consultation. Information Commissioner's Office, Wilmslow, p.94-95. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (accessed 15 May 2020).

[162] Ibid, p.95.

- Consider the system requirements necessary to support a meaningful human review **from the design phase.** Particularly, the interpretability requirements and effective user-interface design to support human reviews and interventions.
- Design and deliver appropriate training and support for human reviewers.
- Give staff the appropriate authority, incentives and support to address or escalate individuals' concerns and, if necessary, override the AI tool's decision.

3.2.3 Selecting non-biased data

Biases are one of the main issues involved in AI development, an issue that contravenes the fairness principle (see "Lawfulness, fairness and transparency principle" within Part II section "Principles" of these Guidelines). Biases might be caused by a lot of different issues. When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. Sometimes, it might happen that datasets are biased due to malicious actions. Feeding malicious data into an AI tool may change its behavior, particularly with self-learning systems.[163] For instance, in the case of chatbot Tay, developed by Microsoft, a huge number of Internet users started posting racists and sexist comments that served to feed the algorithm. As a final result, Tay started sending racist and sexist tweets after just a few hours of operation. On other occasions, the main problem is that the dataset does not represent well the population under consideration and for the intended purpose. Therefore, it contains hidden bias that will be transposed to the trained tool that will reflect such biases, and this might lead to the results of the model being incorrect or discriminatory.[164]

Therefore, issues related to the composition of the databases used for training raise crucial ethical and legal issues, not only issues of efficiency or of a technical nature. Thus, they need to be addressed prior to training the algorithm. The AI models must "be trained using relevant and correct data and it must learn which data to emphasize. The model must not emphasize information relating to racial or ethnic origin, political opinion, religion or belief, trade union membership, genetic status, health status or sexual orientation if this would lead to arbitrary discriminatory processing."[165] Identifiable and discriminatory bias should be removed in the dataset building phase where possible.

Box 18: Understanding biases: the gorilla case

In 2015, a software engineer, Jacky Alciné denounced the image recognition algorithms

[163] High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI. European Commission, Brussels, p.17. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 15 May 2020).

[164] For a definition of direct and indirect discrimination, see, for instance, Article 2 of Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation. See also Article 21 of the Charter of Fundamental Rights of the EU.

[165] Norwegian Data Protection Authority (2018) Artificial intelligence and privacy. Norwegian Data Protection Authority, Oslo. Available at: https://iapp.org/media/pdf/resource_center/ai-and-privacy.pdf (accessed 15 May 2020).

used in Google Photos that classified some black people as "gorillas." Google recognized the issue immediately and promised to fix it.

This glitch was produced by a serious mistake in the training phase. The algorithm was trained to recognize people with a dataset that was primarily composed of photographs of Caucasian people. As a consequence, the algorithm considered that a black person was much more similar with the object "gorilla" that it had been trained to recognize, than with the object "human". This example shows perfectly well the importance of data selection for training purposes.

Thus, to integrate ethical requirements into this phase, the AI developer should evaluate the ethical consequences of data selection in relation to diversity and make changes, if necessary. Indeed, the controller "should use appropriate mathematical or statistical procedures for the profiling, implement technical and organizational measures appropriate to ensure, in particular, that **factors which result in inaccuracies in personal data are corrected and the risk of errors is minimized,** secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect."[166]

Controllers should always keep in mind that what makes this issue so specific is that selecting a dataset for training involves making decisions and choices at times in an almost unconscious manner (whereas coding a traditional, deterministic algorithm is always a deliberate operation). Whoever trains an algorithm in some ways builds into it their own way of seeing the world, values or, at the very least, the values which are more or less directly inherent in the data gathered from the past.[167] This means that **the teams in charge of selecting the data to be integrated in the datasets should be composed of people that ensure the diversity that the AI development is expected to show.** In any case, legal expertise on anti-discrimination regulation might be relevant to this point.

4 Modeling (training)

"In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, several techniques exist for the same data mining problem type. Some techniques have specific requirements on the

[166] Recital 71 of the GDPR.

[167] CNIL (2017) How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence. Commission Nationale de l'Informatique et des Libertés, Paris, p.34. Available at: www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf (accessed 15 May 2020).

form of data. Therefore, stepping back to the data preparation phase may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models."[168]

4.1 Description

This phase involves several key tasks. Overall, the developer must do the following:

- **Select the modeling technique that will be used.** Depending on the type of technique, consequences such as data inference, obscurity or biases are more or less likely to happen.
- **Make a decision on the training tool to be used.** This enables the developer to measure how well the model can predict history before using it to predict the future. Training always involves running empirical testing with personal data. Sometimes, developers test the model with data that are different from those used to generate it. Therefore, at this stage one might talk about different types of datasets. Sometimes identifying the individuals that the training data relates to might be difficult. This creates issues for fulfilling individuals' rights that should be addressed appropriately.

4.2 Main actions that need to be addressed

4.2.1 Implementing data minimization principle

According to the Principle of purpose limitation (see "Purpose limitation principle" within Part II, section "Principles" of these Guidelines), controllers using AI tools determine the purpose of the AI tool's use at the outset of its training or deployment, and perform a re-assessment of this determination should the system's processing throw up unexpected results, since it requires that personal data only be collected for "specified, explicit and legitimate purposes" and not used in a way that is incompatible with the original purpose.

According to the data minimization principle, controllers must proceed to reduce the amount of data and/or the range of information about the data subject they provide as soon as possible. Consequently, personal data used during the training phase have to be purged of all information not strictly necessary for training of the model (see the "Temporal aspect" subsection in the "Data minimization" section of the "Principles" in Part II). There are multiple strategies to ensure data minimization at the training stage.

[168] Shearer, C. (2000) 'The CRISP-DM model: the new blueprint for data mining', *Journal of Data Warehousing* 5(4): 13-23, p.17. Available at: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf (accessed 15 May 2020).

Techniques are continuously evolving. However, some of the most common are given below;[169] see also the "Integrity and confidentiality" section in the "Principles", Part II):

- Analysis of the conditions that the data must fulfil in order to be considered of high quality and with a great predictive capacity for the specific application.
- Critical analysis of the extent of the data typology used in each stage of the AI tool.
- Deletion of unstructured data and unnecessary information collected during the pre-processing of the information.
- Identification and suppression of those categories of data that do not have a significant influence on learning or on the outcome of the inference.
- Suppression of irrelevant conclusions associated with personal information during the training process, for example, in the case of unsupervised training.
- Use of verification techniques that require less data, such as cross-validation.
- Analysis and configuration of algorithmic hyperparameters that could influence the amount or extent of data processed in order to minimize them.
- Use of federated rather than centralized learning models.
- Application of differential privacy strategies.
- Training with encrypted data using homomorphic techniques.
- Data aggregation.
- Anonymization and pseudonymization, not only in data communication, but also in the training data, possible personal data contained in the model and in the processing of inference.

4.2.2 Detecting and erasing biases

Even though the mechanisms against biases are conveniently adopted in previous stages (see previous section about training), it is still necessary to ensure that the results of the training phase minimize biases. This can be difficult, since some types of bias and discrimination are often particularly hard to detect. The team members who are curating the input data are sometimes unaware of them, and the users who are their subjects are not necessarily cognisant of them either. Thus, the monitoring systems implemented by the AI developer in the validation stage are extremely important factors to avoid biases.

There are a lot of technical tools that might serve well to detect biases, such as the Algorithmic Impact Assessment.[170] The AI developer must consider their effective

[169] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.40. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 15 May 2020).

[170] Reisman, D., Crawford, K. and Whittaker, M. (2018) Algorithmic impact assessments: a practical framework for public agency accountability. AI Now Institute, New York, NY. Available at: https://ainowinstitute.org/aiareport2018.pdf (accessed 15 May 2020).

implementation.[171] However, as the literature shows[172], it might happen that an algorithm cannot be totally purged of all different types of biases. The developer should however try to at least be aware of their existence and the implications that this might bring (see "Lawfulness, fairness and transparency" and "Accuracy" sections in "Principles", Part II of these Guidelines).

4.2.3 Exercising data subjects' rights

Quite obviously, controllers must facilitate all data subjects' rights in the whole life cycle. However, in this specific stage, right to access, rectification and erasure are particularly sensitive and include certain characteristics of which controllers need to be aware.

a) Right of access

In general, training data can scarcely be linked to an individual data subject, since they usually only include information relevant to predictions, such as past transactions, demographics, or location, but not contact details or unique customer identifiers. Moreover, they are often pre-processed, to make them more amenable to the algorithms. However, this does not mean that these data can be considered as entirely pseudonymized or anonymized. Thus, they continue to be personal data. For instance, in the case of a purchase prediction model, the training might include a pattern of purchases unique to one customer. In this example, if a customer were to provide a list of their recent purchases as part of their request, the organization may be able to identify the portion of the training data that relates to that individual.

Under such circumstances, AI developers should respond to data subjects' requests to gain access to their personal data, assuming they have taken reasonable measures to verify the identity of the data subject, and no other exceptions apply. And, as the ICO states, "requests for access, rectification or erasure of training data should not be regarded as manifestly unfounded or excessive just because they may be harder to fulfil or the motivation for requesting them may be unclear in comparison to other access requests an organization typically receives." [173]

However, it is clear that organizations do not have to collect or maintain additional personal data to enable identification of data subjects in training data for the sole purposes of complying with the regulation. If the AI developers cannot identify a data subject in the training data and the data subject cannot provide additional information

---

[171] ICO (2020) AI auditing framework - draft guidance for consultation. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf (accessed 15 May 2020).

[172] Chouldechova, A. (2017) 'Fair prediction with disparate impact: a study of bias in recidivism prediction instruments', *Big Data* 5(2): 153-163, http://doi.org/10.1089/big.2016.0047

[173] ICO (2019) Enabling access, erasure, and rectification rights in AI tools. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/ (accessed 15 May 2020).

that would enable their identification, AI developers are not obliged to fulfil a request that is not possible to satisfy.

b) Right to rectification

In the case of the right to rectification, the controller must guarantee the right of rectification of the data, especially those generated by the inferences and profiles drawn up by the AI development.

Even though the purpose of training data is to train models based on general patterns in large datasets and thus individual inaccuracies are less likely to have any direct effect on a data subject, the right to rectification cannot be limited. As a maximum, the controller could ask for a longer period (two extra months) to proceed with the rectification if the technical procedure is particularly complex (Article 11(3)).

Box 19: Rectification

As an example: it may be more important to rectify an incorrectly recorded customer delivery address than to rectify the same incorrect address in training data. This is because the former could result in a failed delivery but the latter would barely affect the overall accuracy of the model.[174]

c) Right to erasure

Data subjects hold a right to request the deletion of their personal data. However, this right might be limited if some concrete circumstances apply. According to the ICO, "organizations may also receive requests for erasure of training data. Organizations must respond to request for erasure when data subjects provided appropriate grounds, unless a relevant legal exemption applies. For example, if the training data is no longer needed because the ML model has already been trained, the organization must fulfil the request. However, in some cases, where the development of the system is ongoing, it may still be necessary to retain training data for the purposes of re-training, refining and evaluating an AI tool. In this case, the organization should take a case-by-case approach to determining whether it can fulfil requests. Complying with a request to delete training data would not entail erasing any ML models based on such data, unless the models themselves contain that data or can be used to infer it."[175]

5 Evaluation (validation)

[174] ICO (2019) Enabling access, erasure, and rectification rights in AI systems. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/ (accessed 15 May 2020).

[175] ICO (2019) Enabling access, erasure, and rectification rights in AI systems. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/ (accessed 15 May 2020).

"Before proceeding to final deployment of the model built by the data analyst, it is important to more thoroughly evaluate the model and review the model's construction to be certain it properly achieves the business objectives. Here it is critical to determine if some important business issue has not been sufficiently considered. At the end of this phase, the project leader then should decide exactly how to use the data mining results. The key steps here are the evaluation of results, the process review, and the determination of next steps."[176]

5.1 Description

This phase involves several tasks that raise important data protection issues. Overall, the developer must:

- evaluate the results of the model, for instance, whether it is accurate or not; to this purpose, the AI developer might test it in the real world
- review the process: the developer could review the data mining engagement to determine if there is any important factor or task that has somehow been overlooked. This includes quality assurance issues.

5.2 Main actions to be addressed

5.2.1 Processes of dynamic validation

The validation of the processing including an AI component must be done in conditions that reflect the actual environment in which the processing is intended to be deployed. Moreover, the validation process requires periodic review if conditions change or if there is a suspicion that the solution itself may be altered. AI developers must make sure that validation reflects the conditions in which the algorithm has been validated accurately.

In order to reach this aim, validation should include all components of an AI tool, including data, pre-trained models, environments and the behavior of the system as a whole and be performed as soon as possible. Overall, it must be ensured that the outputs or actions are consistent with the results of the preceding processes, comparing them to the previously defined policies to ensure that they are not violated.[177] Validation sometimes needs gathering new personal data. In other cases, controllers use data for

[176] Shearer, C. (2000) 'The CRISP-DM model: the new blueprint for data mining', *Journal of Data Warehousing* 5(4): 13-23, p.17. Available at: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf (accessed 15 May 2020).

[177] High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI. European Commission, Brussels, p.22. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 15 May 2020).

purposes others than the original ones. In all these cases, controllers should ensure compliance with the GDPR (see the "Purpose limitation" section in the "Principles" and "Data protection and scientific research" in the "Main Concepts" section, both within Part II).

5.2.2 Deleting unnecessary dataset

Quite often, the validation and training processes are somehow linked. If the validation recommends improvements in the model, training should be performed again.

In principle, once the AI development has finally been achieved, the training stage of the AI tool is completed. At that moment, you should implement the removal of the dataset used for this purpose, unless there is a lawful need to maintain it for the purpose of refining or evaluating the system, or for other purposes compatible with those for which they were collected in accordance with the conditions of Article 6(4) of the GDPR (see "Define adequate data storage policies" section). However, AI developers should always consider that deleting the personal data can work against the need to update the accuracy of tools based on the real-time self-learning of algorithms: if a mistake is found, they will probably need to recall the data previously used in the training stage.

In the event that data subjects request its deletion, the controller shall have to adopt a case-by-case approach taking into account any limitations to this right provided by the Regulation (see Article 17(3)).[178]

5.2.3 Performing external audit of data processing

In cases where the risks of the processing of personal data within the AI tool are high, **an audit of the system by an independent third party must be considered**. A variety of different audits can be used. These might be internal or external, they might cover the final product only, or be performed with less evolved prototypes. They could be considered a form of monitoring or a transparency tool. Annex I, at the end of this document, contains some recommendations by the Spanish Data Protection Agency that could serve as a model.

In terms of legal accuracy, AI tools must be audited to verify whether processing of personal data within their system fulfil obligations stipulated in GDPR considering wide range of issues that are arousing. The High-Level Expert Group on AI stated that "testing processes should be designed and performed by as diverse group of people as possible. Multiple metrics should be developed to cover the categories that are being tested for different perspectives. Adversarial testing by trusted and diverse "red teams" deliberately attempting to "break" the system to find vulnerabilities, and "bug bounties" that incentivize outsiders to detect and responsibly report system errors and weaknesses,

[178] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.26. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 15 May 2020).

can be considered."[179] However, there are good reasons to be sceptical about the capability of an auditor to check the functioning of a machine learning system.

This is why it is sensible to focus on the items included by the AEPD in its recommended checklist: it would be more straightforward to focus on the measures implemented to avoid biases, obscurity, hidden profiling, etc., focusing on the implementation of principles such as data protection by design and by default (see "Data protection by design and by default" within Part II section "Main concepts" of these Guidelines) or data minimization (see "Data minimization principle" within Part II section "Principles" of these Guidelines and the adequate use of tools such as the DPIA or the intervention of a skilled DPO, than trying to have an in depth understanding of the functioning of a complex algorithm (the "black box" problem is obviously very important to this purpose). Implementing adequate data protection policies form the first stages of the lifecycle of the tool is the best way to avoid data protection issues.

Box 20: The difficulty in auditing a machine-learning system: IBM's Watson platform

IBM's policy stresses that Watson is trained via "supervised learning". In other words, the system is guided, step-by-step, in its learning. This should mean its process can be monitored, as opposed to unsupervised learning, in which the machine has full autonomy in determining its operating criteria. IBM also claims to check what the systems have been doing, before any decision to retain a certain type of learning. But experts researching this subject who have spoken out during the various organized debates (not least by Allistene's research committee on ethics, CERNA) have insisted time and again that such statements are erroneous. Based on current research, the "output" produced by the most recent machine learning algorithms is not explainable, explainable AI being a concept on which research is ongoing. They also point out that it is very difficult to audit a machine learning system in practice.[180]

6 Deployment

"Deployment is the process of getting an IT system to be operational in its environment, including installation, configuration, running, testing, and making necessary changes. Deployment is usually not done by the developers of a system but by the IT team of the

[179] High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI. European Commission, Brussels, p.22. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 15 May 2020).

[180] CNIL (2017) How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence. Commission Nationale de l'Informatique et des Libertés, Paris, p.28. Available at: www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf (accessed 15 May 2020).

customer. Nevertheless, even if this is the case, developers will have a responsibility to supply the customer with sufficient information for successful deployment of the model. This will normally include a (generic) deployment plan, with necessary steps for successful deployment and how to perform them, and a (generic) monitoring and maintenance plan for maintenance of the system, and for monitoring the deployment and correct usage of data mining results."[181]

6.1 Main actions that need to be addressed

6.1.1 General remarks

At the time of distribution of the AI tool, if it incorporates personal data, it will be necessary to do the following (see also the "Purchasing or promoting access to a database" section in the "Main Tools and Actions" in Part II):

- To delete them or, on the contrary, to justify the impossibility of doing so, completely or partly because of the degradation it would mean for the model (see the "Storage limitation" section in the "Principles" in Part II).
- Determine the legal basis for carrying out the communication of personal data to third parties, especially if special categories of data are involved (see the "Lawfulness" subsection in the "Lawfulness, fairness and transparency" section in Part II of these Guidelines).
- Inform the data subjects of the processing above.
- Demonstrate that the data protection by design and by default policies have been implemented (especially data minimization).
- Depending on the risks it might pose to the stakeholders and taking into account the volume or categories of personal data to be used, conducting a Data Protection Impact Assessment (DPIA) should be considered.[182] (see "DPIA" within Part II section "Main actions and tools" of these Guidelines)

In principle, once the model is put into use, the training data is removed from the algorithm and the model will only process the personal data to which it is applied. The controller might retain the data subject's data for the customization of the service being offered by the AI tool. However, once this service finishes, these data must be deleted, unless convincing reasons make it recommendable to keep them. This does not mean that data storage should last forever, of course.

The AI developer must make sure that the algorithm does not include personal data in a hidden way (or take necessary measures if this is unavoidable). In any case, the

[181] SHERPA project (2020) Guidelines for the ethical development of AI and big data systems: an ethics by design approach. SHERPA, p.13. Available at: www.project-sherpa.eu/wp-content/uploads/2019/12/development-final.pdf (accessed 15 May 2020).

[182] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.26. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 15 May 2020).

developer must perform a formal evaluation assessing which personal data from the data subjects could be identifiable.[183] This can be complicated at times. For example, some AI tools, such as Vector Support Machines (VSM) could contain examples of the training data by design within the logic of the model. In other cases, patterns may be found in the model that identify a unique individual.[184] In all of these cases, unauthorized parties may be able to recover elements of the training data, or infer who was in it, by analyzing the way the model behaves.

Under such conditions, it might be difficult to ensure that the data subjects are able to exercise and fulfil their rights of access, rectification, and erasure (see "Right to access, rectification and erasure" sections within Part II section "Data subject's rights" of these Guidelines. Indeed, "unless the data subject presents evidence that their personal data could be inferred from the model, the controller may not be able to determine whether personal data can be inferred and therefore whether the request has any basis."[185] However, controllers should take regular action to proactively evaluate the likelihood of the possibility of personal data being inferred from models in light of the state-of-the-art technology, so that the risk of accidental disclosure is minimized. If these actions reveal a substantial possibility of data disclosure, necessary measures to avoid it should be implemented (see the "Integrity and confidentiality" section in the "Principles", Part II of these Guidelines).

6.1.2 Updating information

If the algorithm is implemented by a third party, the AI developers should communicate the results of the validation and monitoring system employed and offer their collaboration to continue monitoring the validation of the results. It would also be advisable to establish this kind of coordination with third parties from whom they acquire databases or any other relevant component in the life cycle of the system. If this involves data processing by a third party, the controller must ensure that access is provided within a legal basis.

It is necessary to offer real time information to the end user about the values of accuracy and/or quality of the inferred information at each stage (see the "Accuracy principle" section in the "Principles" in Part II). When the inferred information does not reach minimum quality thresholds, it must be explicitly indicated that this information has no value.[186] This requirement often implies that developers should provide detailed

---

[183] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.41. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 15 May 2020).

[184] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.13. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 15 May 2020).

[185] ICO (2019) Enabling access, erasure, and rectification rights in AI tools. Information Commissioner's Office, Wilmslow. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/ (accessed 15 May 2020).

[186] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p.35. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 15 May 2020).

information about the training and validation stages. Information about the datasets used for those purposes is particularly important. Otherwise, the use of the solution might bring disappointing results to the end users, who are left speculating on the cause.

Annex I: Auditing AI tools[187]

According to the Spanish Data Protection Agency, audits should cover a large list of items, namely:

- The existence or not of personal data, profiling or automatic decisions on data subjects without human intervention.
- The effectiveness of anonymization and pseudonymization methods.
- The existence and legitimacy of the processing of special categories of data, in particular the inferred information.
- The legal basis for the processing and the identification of responsibilities.
- In particular, where the legal basis is the legitimate interest, assessment of the balance between the various interests and the impacts on rights and freedoms of the data subjects in the light of the guarantees adopted.
- The information and the effectiveness of the transparency mechanisms implemented.
- The application of the principle of proactive accountability and risk management for the rights and freedoms of the data subjects and in particular, whether the obligation or need for the execution of the DPIAs and, if so, their results.
- The application of data protection measures by design and by default, such as:
    - the analysis of the need of the quantity and extension of personal data processing in the different stages of the AI development;
    - the analysis of the accuracy, reliability, quality and biases of the data used or captured for the development or operation of the AI component, as well as the data cleansing methods used;
    - the monitoring and implementation of testing and validation processes concerning the precision, accuracy, convergence, consistency, predictability and any another metric of the goodness of the algorithms used, profiled and the inferences made. In addition, checking that these parameters meet the processing requirements.
- The adequacy of security measures to avoid risks to privacy.
- The training and education of the staff of the controller linked to the development or implementation of the IAI component, where appropriate, in the latter case with particular attention to the correct interpretation of the inferences.
- The need and, where appropriate, the capacity of the DPO.
- The incorporation of mechanisms to ensure attention to the rights of data subjects, in particular the ex officio deletion of personal data, with special attention to the rights of minors.
- The compliance with the limitations on automatic decisions without human

[187] AEPD (2020) Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción. Agencia Espanola Proteccion Datos, Madrid, p. 45-47. Available at: www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf (accessed 3 June 2020).

Annex II. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness[188]

Inception

1. What is the health question relating to patient benefit?

2. What evidence is there that the development of the algorithm was informed by best practices in clinical research and epidemiological study design?

Study

1. When and how should patients be involved in data collection, analysis, deployment, and use?

2. Are the data suitable to answer the clinical question—that is, do they capture the relevant real world heterogeneity, and are they of sufficient detail and quality?

3. Does the validation methodology reflect the real-world constraints and operational procedures associated with data collection and storage?

4. What computational and software resources are required for the task, and are the available resources sufficient to tackle this problem?

Statistical methods

1. Are the reported performance metrics relevant for the clinical context in which the model will be used?

2. Is the ML/AI algorithm compared to the current best technology, and against other appropriate baselines?

3. Is the reported gain in statistical performance with the ML/AI algorithm justified in the context of any trade-offs?

Reproducibility

1. On what basis are data accessible to other researchers?

2. Are the code, software, and all other relevant parts of the prediction modeling pipeline available to others to facilitate replicability?

3. Is there organizational transparency about the flow of data and results?

Impact evaluation

1. Are the results generalizable to settings beyond where the system was developed (that is, results reproducibility/external validity)?

[188] Vollmer, S. et al. (2020) 'Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness', *BMJ* 2020;368:l6927, http://dx.doi.org/10.1136/bmj.l6927

2. Does the model create or exacerbate inequities in healthcare by age, sex, ethnicity, or other protected characteristics?

3. What evidence is there that clinicians and patients find the model and its output (reasonably) interpretable?

4. How will evidence of real-world model effectiveness in the proposed clinical setting be generated, and how will unintended consequences be prevented?

Implementation

1. How is the model being regularly reassessed, and updated as data quality and clinical practice changes (that is, post-deployment monitoring)?

2. Is the ML/AI model cost effective to build, implement, and maintain?

3. How will the potential financial benefits be distributed if the ML/AI model is commercialized?

4. How have the regulatory requirements for accreditation/approval been addressed?

Annex III: Checklists

Checklist: business understanding

 The controllers have assessed the amount of data that will be needed to develop the AI solution or the nature of that data and ensured that they work well with the minimization principle.

 The controllers have fixed acceptable thresholds of false positives/negatives or ranges, depending on the use case and then have performed a utility balance.

 The controllers have adequately balanced the level of accuracy needed and the range of personal data required to reach it.

 The controllers have provided for the development of more understandable algorithms over less understandable ones whenever possible.

 The controllers have ensured an optimal training for all subjects involved in the project or an adequate internal or external assessment on ethical and legal issues.

 The controllers have carefully designed the tools that will legitimate data processing. To this purpose, they have checked if the intervention of an ethics committee is needed or whether any kind of soft regulation is applicable.

 The controllers have adopted a risk-based approach (including technical and organizational security measures) that minimizes the risks to data subjects' rights, interests, and freedoms.

 The controllers have implemented tools and policies aimed at assessing and evaluating the effectiveness of technical and organizational measures regularly.

- The controllers have considered whether the regulatory framework regarding scientific research applies.
- The storage policies keep personal data in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed.
- The controllers have considered the appointment of a DPO

Checklist: data understanding

- The controllers have implemented appropriate technical and organizational measures for ensuring that, by default, only personal data that are necessary for each specific purpose of the processing are processed.
- The controllers have introduced policies that minimize the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. Such measures ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.
- The controllers do not to collect unnecessary data. If data is already stored, they have taken actions aimed at deleting unnecessary data elements.
- The controllers have limited the resolution of the data to what is minimally necessary for the purposes pursued by the processing.
- The controllers have selected the legal basis that most closely reflects the true nature of their relationship with the individual and the purpose of the processing.
- The controllers have carefully analyzed whether processing involves de-anonymizing anonymized data and creating new personal information that was not contained in the original data set and take adequate measures to face these challenges.
- The controllers have made sure that merging datasets does not create ethical or legal issues regarding data subjects' rights and freedoms.

Checklist: Data preparation

- The controllers have ensured that data are accurate, that is, correct and up to date data.
- If profiling or automated decision-making is foreseen:
  - The controllers have sent individuals a link to their privacy statement when they have obtained their personal data indirectly.
  - The controllers have explained how people can access details of the information that they used to create their profile.
  - The controllers have communicated the data subjects who provide them with their personal data and how they can object to profiling.
  - The controllers have introduced procedures for customers to access the personal data input into their profiles, so they can review and edit for any accuracy issues.
  - The controllers have implemented additional checks in place for their profiling/automated decision-making systems to protect any vulnerable groups (including children).
  - The controllers have ensured that they only collect the minimum amount of data needed and have a clear retention policy for the profiles that they create.
  - The controllers have carried out a DPIA to consider and address the risks when they start any new automated decision-making or profiling.
  - The controllers have involved the corresponding DPO in these activities.
  - The controllers have considered the system requirements necessary to support a meaningful human review **from the design phase.** Particularly, the interpretability requirements and effective user-interface design to support human reviews and interventions.
  - The controllers have designed and delivered appropriate training and support for human reviewers.
  - The controllers have given the staff involved in the processing the appropriate authority, incentives and support to address or escalate individuals' concerns and, if necessary, override the AI system's decision.
- The controllers have ensured that the teams in charge of selecting the data to be integrated in the datasets are composed of people that ensure the diversity that the AI development is expected to show.
- The controllers have ensured that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimized.
- The controllers have implemented tools aimed at preventing discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect.

Checklist: Modeling (training)

□ The controllers have determined the purpose of the AI system's use at the outset of its training or deployment, and performed a re-assessment of this determination if the system's processing threw up unexpected results.

□ The controllers have purged the data used during the training phase of all information not strictly necessary for training of the model.

□ The controllers have considered implementing technical tools that might serve well to detect biases, such as the Algorithmic Impact Assessment.

□ The controllers have considered conducting a DPIA at this stage.

□ The controllers have ensured that they are able to respond to data subjects' requests to exceptions to the right to access apply.

□ The controllers can guarantee the right of rectification of the data, especially those generated by the inferences and profiles drawn up by the AI development.

□ The controllers are able to respond to requests for erasure, unless a relevant exemption applies and provided the data subject has appropriate grounds.

Checklist: evaluation (validation)

□ The controllers have made sure that validation reflects the conditions in which the algorithm has been validated accurately.

□ The controllers have informed data subjects about additional processing at this stage.

□ The controllers have ensured the removal of the dataset used for validation purposes, unless there is a lawful need to maintain them for the purpose of refining or evaluating the system, or for other purposes compatible with those for which they were collected.

□ The controllers have considered conducting a DPIA at this stage.

□ If the data subjects request the deletion of their data, the controller have adopted a case-by-case approach taking into account any limitations to this right provided by the Regulation).

□ The controllers have considered an audit of the system by an independent third party.

Checklist: deployment

□ The controllers have deleted all unnecessary personal data or, on the contrary, justified the impossibility of doing so.

□ The controllers have informed data subjects about additional processing at this stage.

□ The controllers have determined the adequate legal basis for carrying out the communication of personal data to third parties, especially if special categories of data are involved.

□ The controllers have considered conducting a DPIA.

☒ The controllers have made sure that the algorithm does not include personal data in a hidden way (or taken necessary measures if this is unavoidable).

☒ The AI developers have implemented tools aimed at communicating the results of the validation and monitoring system employed and offered their collaboration to continue.

☐ The AI developers have a commitment to offer real time information to the end users about the values of accuracy and/or quality of the inferred information at each stage.

First scenario: building and AI tool devoted to diagnosing COVID-19 disease

Iñigo de Miguel Beriain (UPV/EHU)

This part of The Guidelines has been reviewed and validated by Marko Sijan, Senior Advisor Specialist, (HR DPA)

Description

The response to the pandemic created situations in which many patients needed health care but it was hard to provide due to the high incidence of the disease amongst health personnel. In this situation, a radiologist, for example, could not deal with the high number of X-Rays to be analyzed due to the absence of colleagues on sick leave. Using AI for such purposes could be of great help for the future, but there are a lot of ethical and legal issues that must be considered. In this scenario, we will analyze the different steps that must be fulfilled by a team of researchers willing to train an algorithm able to help with the diagnosis of lung disease.

Preliminary remarks

Research with health data presents particularly important ethical challenges. If we are also talking about a case where patients suffer from a disease such as COVID, the dilemma is particularly pressing. In the context of health care, it is easy to intermix informed consent associated with clinical practice with consent to biomedical research. This is always a matter of concern. The two things are hugely different. Planning an activity such as the development of an IA tool for diagnosis should take this into account. This is especially true for patients in more vulnerable situations than usual. It should never be forgotten that the objectives of biomedical research cannot overlap with people's interests and well-being.

There are several essential tools that researchers should always keep in mind when designing a plan for the development of an AI tool. The Ethics issues checklist included in the Horizon 2020 Programme Guidance "How to complete your ethics self-assessment", page 6[189] is highly recommended. Some essential documents to be consulted include:

- High-Level Expert Group on AI: 'Ethics guidelines for trustworthy AI'.[190]
- EU Commission, White Paper on Artificial Intelligence - A European approach to excellence and trust[191]
- **T**raining and **R**esources in **R**esearch **E**thics **E**valuation (TRREE)[192] is an online tool that provides free-of-charge access to:
    - **e-Learning:** a distance learning program and certification on research ethics evaluation,
    - **e-Resources:** a participatory web-site with international, regional and national regulatory and policy resources.
- Additional online training tools can be found in the EUREC web page[193].

Step by Step Analysis

1 Business understanding

1.1 Description

189

ttps://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-self-assess_en.pdf

190 https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

191 https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

192 https://elearning.trree.org/mod/page/view.php?id=70

193 http://www.eurecnet.org/materials/index.html

"The initial business understanding phase focuses on understanding the project objectives from a business perspective, converting this knowledge into a data mining problem definition, and then developing a preliminary plan designed to achieve the objectives. In order to understand which data should later be analyzed, and how, it is vital for data mining practitioners to fully understand the business for which they are finding a solution. The business understanding phase involves several key steps, including determining business objectives, assessing the situation, determining the data mining goals, and producing the project plan."[194]

This general objective involves four main tasks:

1. Determine the Business Objectives. This means:
    a. Uncover the primary business objective as well as the related questions the business would like to address.
    b. Determine the measure of success.
2. Assess the Situation
    a. Identify the resources available to the project, both, material and personal.
    b. Identify what data is available to meet the primary business goal.
    c. List the assumptions made in the project.
    d. List the project risks, list potential solutions to those risks, create a glossary of business and data mining terms, and construct a cost-benefit analysis for the project.
3. Determine the Data Mining Goals: decide what level of predictive accuracy is expected to consider the project successful.
4. Produce a Project Plan: describe the intended plan for achieving the data mining goals, including outlining specific steps and a proposed timeline, an assessment of potential risks, and an initial assessment of the tools and techniques needed to support the project.

1.2 Main actions that need to be addressed

1.2.1 Defining business objectives

The first thing to clarify when you want to create an AI tool is what you want to achieve. In the case of a tool that identifies a pathology from an X-ray, it may be, for example, that

1) It is meant to serve as a support for the radiologist's work
2) It may be used to support the work of a primary care physician, that is, to determine whether to refer the patient to a specialist.
3) It can also be designed to replace the physician and make a diagnosis of, for example, COVID on its own.
4) It can be used for performing a first triage (this is, recommending whether a primary care physician or a specialist should intervene).

[194] Shearer, Colin, The CRISP-DM Model: The New Blueprint for Data Mining, p. 14.

Each of these scenarios has vastly different characteristics. Some of them require a higher level of accuracy than others. Thus, for example, if you intend to replace the health professional, it is necessary for the AI to reach an impressively high level of precision.

The ethical and legal implications of the different purposes are, at the same time, vastly different. If the mechanism is to be used for automated decision-making purposes, as in cases 3) or 4), the processing of the data will be subject to a considerably stricter legal regime. In fact, in many countries such use may be directly illegal.

All these considerations must be borne in mind from the outset. The development process should not be initiated if you, as the controller, do not clarify what results are to be achieved, because this issue is key in determining whether or not the planned data processing is in line with GDPR. Deciding the level of predictive accuracy expected to consider the project successful, is essential to assess the amount of data that will be needed to develop the AI tool or the nature of that data. The level of predictability or precision of the algorithm, the validation criteria to test it, the maximum quantity or the minimum quality of the personal data that will be necessary to use it in the real world, etc., are fundamental features of an AI development.

These key development elements should be considered from the first stage of the solution's life cycle. This will be extremely helpful to implement a data protection by design policy (see “Data protection by design and by default” within Part II section “Main concepts” of these Guidelines). If an acceptable level of accuracy could be reached by using considerably less amount of personal data than a higher level requires, this should be strongly considered. The more inaccurate you are about these assessments, the more difficult it becomes to determine the precise purposes that are pursued by processing (see “Prerequisites to lawfulness specified, explicit purposes” subsection in “Lawfulness, fairness and transparency” section in “Principles” in Part II). If we keep in mind that controllers must make the purposes of processing explicit, that is, “revealed, explained or expressed in some intelligible way”, accurate expectations are strongly recommendable.

1.2.2 Opting for the technical solutions

In general, you should always provide for the development of more understandable algorithms over less understandable ones. Trade-offs between the explainability/transparency and best performance of the system must be appropriately balanced based on the context of use. Even though in healthcare the accuracy and performance of the system may be more important than its explainability, you should always keep in mind that explaining a recommendation could serve well to train doctors, provide adequate information to patients who have to make a choice between different possible treatments or to justify a triage decision, for instance. Thus, if a quite similar service can be offered either through an easy to understand algorithm or an opaque one, that is, when there is no trade-off between explainability and performance, you should opt for the one that is more interpretable (see “Lawfulness, fairness and transparency” section in “Principles” in Part II).

1.2.3 Implementing a training program on ethical and legal issues

This action is one of the most important pieces of advice to be considered from the very first moment of an AI business development. Algorithm designers (developers, programers, coders, data scientists, engineers), who occupy the first link in the algorithmic chain, are likely to be unaware of the ethical and legal implications of their actions. If all intervening staff are in close contact with the data subjects, ethical considerations are easier to implement. However, this will probably not be your case. Indeed, one of the main problems that an AI tool devoted to dealing with health care issues is that it generally uses personal data that are included in large datasets. This somehow blurs the relationship between the data and the data subject, leading to violations of the regulations that rarely occur when the controller and the subject have a direct relationship.

This could bring terrible consequences in terms of adequate compliance with data protection standards, mainly since data of special categories are at stake. It is paramount that these key workers have the fullest possible awareness of the ethical and social implications of their work, and of the very fact that these can even extend to societal choices, which they should not by rights be able to judge alone. Silo mentality must be carefully fought.

In order to avoid that the misrepresentation of the ethical and legal issues provokes unwanted consequences, there are two main courses of action that can be adopted. First, developers might try to ensure that algorithm designers are able to understand the implications of their actions, both for individuals and society, and be aware of their responsibilities by learning to show continued attention and vigilance.[195] In that sense, an optimal training for all subjects involved in the project (developers, programers, coders, data scientists, engineers, researchers) even before it starts could be one of the most efficient tools to save time and resources in term of compliance with data protection regulation. Thus, implementing basic training programs that include at least the fundamentals of the Charter of Fundamental Rights, the principles exposed in Article 5 of the GDPR, the need for a legal basis for processing (including contracts between the parties), etc.

However, training people who have never been in touch with data protection issues might be hard. An alternative policy is the involvement of an expert on data protection, ethical and legal issues in the development team, so as to create an interdisciplinary team. This might be done by hiring an expert for this purpose (an internal worker or an external consultant) to design the strategy and the subsequent decisions on personal data required by the development of the tools, with the close involvement of the Data Protection Officer.

Adopting adequate measures in terms of ensuring confidentiality is also strongly recommendable (see “Measures in support of confidentiality” subsections in the “Integrity and confidentiality” section in “Principles” within Part II of these Guidelines).

---

[195] Ibid., p.55.

1.2.4 Designing legitimate data processing tools

According to article 5(1)(a) of the GDPR, personal data shall be "collected for specific, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes". The concept of legitimacy is not well defined in the GDPR, but the Article 29 Working Party stated that legitimacy involves that data must be processes "in accordance with the law", and "law" should be understood as a broad concept that includes "all forms of written and common law, primary and secondary legislation, municipal decrees, judicial precedents, constitutional principles, fundamental rights, other legal principles, as well as jurisprudence, as such 'law' would be interpreted and taken into account by competent court".

Therefore, it is a wider concept than lawfulness. It involves compliance with the main values of the applicable regulation and the main ethical principles at stake. For instance, some concrete AI developments will need the intervention of an ethics committee. In other cases, guidelines or any other kind of soft regulation might be applicable. You should ensure adequate compliance with this requirement by designing a plan form this preliminary stage of the lifecycle of the tool (see "Legitimacy and lawfulness" in "Lawfulness, fairness and transparency" of the "Principles" in Part II). To this purpose, you should be particularly aware of the requirements posed by the applicable regulation at the national level. In many Member states, developing an algorithm related to health care will surely involve the intervention of Ethics Committees, most probably at a preliminary stage. Make sure that your research plan fits well with such requirements.

1.2.5 Adopting a risk-based thinking approach

Since the creation of your algorithm will surely involve the use of a huge amount of special categories of personal data, mainly health data, you must ensure that you implement appropriate measures to minimize the risks to data subjects' rights and freedoms (see "Integrity and confidentiality" of the "Principles" in Part II). To this purpose, you must assess the risks to the rights and freedoms of individuals participating in the research and development process and judge what is appropriate to protect them. In all cases, you need to ensure that they comply with data protection requirements.

Risk-based thinking with regard to confidentiality of data, or a risk-based approach to questions of what harm may be done to people/data subjects, must be included from the first steps of the process. It might have legal consequences for the data controller in relation with the obligations stipulated in the GDPR if it is only considered later. Thus, you must identify the implicit threats to the planned data processing and assess the level of intrinsic risk involved. If you are planning to use software for processing purposes, you should ensure that adequate measures in support of confidentiality are implemented. If your AI will use third party software or off-the-shelf software, it is vital that functions that process personal data that have no legal basis, or are not compatible with the intended purposes, are excluded.

Whenever possible, try to avoid using data storage or software services that are located in a third country. If this is unavoidable, you must ensure that your data processing contracts with those third parties provide adequate GDPR compliant protection or, if this is not the case, ensure that the research participants are fully aware of the

privacy/security risks to their data. *You should also be aware and informed about appropriate security measures implemented by data storage and* software service providers, and that the omissions in security may result in a breach of security processing.

In addition, you must ensure that appropriate technical and organizational measures are implemented to eliminate, or at least mitigate the risk, reducing the probability that the identified threats will materialize or reducing their impact. The security measures must become a part of your records of processing (see "Documentation of Processing" section in "Main Tools and Actions" within Part II of these Guidelines) and all implemented measures will be part of the DPIA **(**see "DPIA" within Part II section "Main actions and tools" of these Guidelines).

Once the selected measures are implemented, the remaining residual risk should be assessed and kept under control. Both the risk analysis and the DPIA are the tools that apply. In your concrete case, you must carry out a DPIA, since the creation of the AI tool will involve the processing on a large scale of special categories of data.

Finally, do not forget that when using big data and AI it is hard to foresee what the future risks will be, so doing assessment of ethical implications will not be sufficient to address all possible risks. Therefore, it is important to consider having a reassessment of risks and also highly recommendable to integrate a more dynamic way of assessing research risks. Do not hesitate to perform additional DPIAs in other stages of the process if need be.

1.2.6 Preparing the documentation of processing

Whoever processes personal data (including both, controllers and processors) needs to document their activities primarily for the use of qualified/relevant Supervisory Authorities. You must do this through records of processing that are maintained centrally by your organization across all its processing activities, and additional documentation that pertains to an individual data processing activity (see "Documentation of Processing" section in "Main Tools and Actions" within Part II of these Guidelines). This preliminary stage is the perfect moment to set up a systematic way of collecting the necessary documentation, since it will be the time when you can conceive and plan the processing activity.

Indeed, you should create a Data Protection Policy (see "Economy of scale for compliance and its demonstration" subsection in "Accountability" section of the "Principles" in Part II) that allows the traceability of information (if approved codes of conduct exist, these should be implemented, again, see "Economy of scale for compliance and its demonstration" subsection in Accountability section of the "Principles" in Part II). This Policy should also make the responsibilities assigned to processors clear, if you are willing to involve them in your project, and include the processing agreement tasks that will be delegated to it in relation to the execution of data subjects' rights. You should always remember that Art. 32(4) GDPR clarifies that an important element of security is to ensure that employees act only on instruction and as instructed by you (see "Integrity and Confidentiality" section in "Principles", Part II of these Guidelines).

The development of your AI tool might involve the use of different datasets. The traceability of the processing, the information about possible re-use of data, and the use of data pertaining to different datasets in different or in the same stages of the life cycle must be ensured by the records.

As stated in the Requirements and acceptance tests for the purchase and/or development of the employed software, hardware, and infrastructure (subsection of the "Documentation of Processing" section, the risk evaluation and the decisions taken "have to be documented in order to comply with the requirement of data protection by design (see "Data protection by design and by default" within Part II section "Main concepts" of these Guidelines). Practically, this can take the form of:

- Data protection **requirements** specified for the purchase (e.g., a tender) or development of software, hardware and infrastructure,
- **Acceptance tests** that verify that the chosen software, systems and infrastructure are fit for purpose and provide adequate protection and safeguards.

Such documentation can be an integral part of the DPIA."

Finally, you should always be aware that, according to Art. 32(1)(d) of the GDPR, data protection is a process. Therefore, **you should test, assess, and evaluate the effectiveness of technical and organizational measures regularly.** This is an excellent moment to build a strategy aimed at facing these challenges.

1.2.7 Regulatory framework usage

The GDPR includes a specific regulatory framework regarding processing for the purposes of scientific research (see "Data protection and scientific research" section in the "Main Concepts" in Part II).[196] Your AI development constitutes scientific research, irrespective of whether it is created for profit or not. Therefore, the "Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfillment of those purposes" (Art. 89(2)). Furthermore, according to article 5 (b) "further processing of the data gathered, in accordance with Article 89(1), would not be considered to be incompatible with the initial purposes ('purpose limitation'). Some other particular exceptions to the general framework applicable to processing for research purposes (such as storage limitation) should also be considered".

You certainly might profit from this favorable framework. Nevertheless, you must be aware of the concrete regulatory framework that applies to this research (mainly, the safeguards to be implemented). It might include important changes depending on respective national regulations. Consultation with your DPO is highly recommended for this purpose.

[196] This specific framework also includes historical research purposes or statistical purposes. However, ICT research is not usually related to these purposes. Therefore, we will not analyze them here.

1.2.8 Defining data storage adequate policies

According to Article 5(1)(e) GDPR, personal data should be "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed" (see "Storage limitation" section of the "Principles" in Part II). This requisite is twofold. On one hand, it relates to identification: data should be stored in a form which permits identification of data subjects for no longer than necessary. Consequently, you should implement policies devoted to avoiding identification as soon as it is not necessary for processing. This involves the adoption of adequate measures to ensure that at any moment, only **the minimal degree of identification that is necessary to fulfill the purposes must be used** (see "Temporal aspect" subsection in "Storage limitation" section of the "Principles" in Part II).

On the other hand, data storage implies that data can only be stored for a **limited period**: the time that is strictly necessary for the purposes for which the data are processed. However, the GDPR permits 'storage for longer periods' if the sole purpose is scientific research (as in your concrete case).

Thus, this exception raises the risk that you decide to keep the data longer than strictly needed so as to ensure that they are available for reasons other than the original purposes they were collected for. Do not do it, if there are no good reasons that recommend it (for instance, if X-Rays come from a medical record, you must keep them in the clinical record of the patient). You must be aware that even though the GDPR might allow storage for longer periods, **you should have a good reason to opt for such an extended period.** Thus, if you do not need the data, and there are no compulsory legal reasons that oblige you to conserve the data, it is better to anonymize or delete it. This could also be an excellent moment to **envisage time limits for erasure of the different categories of data and document these decisions** (see "Accountability Principle" within "Principles" in Part II).

1.2.9 Appointing a Data Protection Officer

According to article 37 GDPR, you must appoint a DPO since you will process a large scale of special categories of data pursuant to Article 9. In any case, key personnel within data controller should elaborate the role of the DPO in relation to the overall management of the project, ensuring that the role of the DPO is not marginal, but cemented into decision-making processes of the organization/project. They should also make clear what that role could be in terms of oversight, decision-making and similar.

1.2.10 Ensuring compliance with legal framework for medical devices

Even though these Guidelines are mainly oriented to data protection issues, we cannot avoid mentioning that you should be well aware from this preliminary stage that you must ensure adequate compliance with the legal framework related to medical devices. We are mainly referring to Regulation (EU) 2017/745 – Medical Devices Regulation (MDR) and Regulation (EU) 2017/746 – In Vitro Diagnostic Medical Devices Regulation (IVDR). Most probably, there will be national regulations applicable to these issues. Please, take actions aimed at compliance. You can find helpful Guidelines to this purpose here: https://ec.europa.eu/docsroom/documents/40323

Regarding the regulation of health data at the Member State Level, this resource might be particularly relevant:

https://ec.europa.eu/health/sites/health/files/ehealth/docs/ms_rules_health-data_en.pdf

2 Data Understanding

2.1 Description

"The data understanding phase starts with an initial data collection. The analyst then proceeds to increase familiarity with the data, to identify data quality problems, to discover initial insights into the data, or to detect interesting subsets to form hypotheses about hidden information. The data understanding phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality".[197]

At this stage, initial data collection takes palace, and an initial study of the data is performed. It involves four sequential tasks:

- Collect initial data
- Describe data
- Analyze data
- Verify data quality.

All of these tasks are aimed at identifying the data available. At this stage, you need to be aware of the data you will have to work with and start making decisions on the way in which main principles related to data protection will be implemented.

2.2 Main actions that need to be addressed

At this stage, there are a huge number of fundamental issues related to the protection of personal data that need to be addressed. Depending on the decisions made, principles such as data minimization, privacy by design or by default, lawfulness, fairness and transparency, etc. will be adequately settled.

2.2.1 Type of collected data

According to the GDPR, you "shall implement appropriate technical and organizational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and

[197] Colin Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, p. 15

their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons."[198] (see "Data Protection by Design and by Default" in Part II, section "Main Concepts") This must be specially kept in mind during this stage, since decisions about the type of data that will be used are often taken at this moment. In general, the simplest way to build your AI in terms of data protection issues would exclusively involve the use of X-Ray images. Nonetheless, it might also be interesting to introduce data related to previous pathologies, age, or gender, for instance. Additionally, one could think about using data such as food habits, zip/postal code, sporting habits, etc. It might happen that adding a lot of new features to the model increases its accuracy in a significant way. However, it could also be possible that this does not happen**. You should balance whether the introduction of additional data apart from the radiographic images, for instance, provides diagnosis with a sufficient level of increased accuracy to justify their use.** This might be difficult to assess in advance, but at least the training phase should clarify this issue. If the increase of accuracy does not justify a disproportionate use of personal data, it should be avoided.

Thus, make sure that you really need huge amounts of data. Smart data might be much more useful than big data. Of course, using smart, well prepared data might involve a huge effort in terms of unification, homogenization, etc., but it will help to implement the principle of data minimization (see "Data minimization principle" within Part II section "Principles" of these Guidelines) in a much more efficient way. To this purpose**, having an expert able to select relevant features might be extremely important.**

Furthermore, you should try to **limit the resolution of the data** to what is minimally necessary for the purposes pursued by the processing. You should also **determine an optimal level of data aggregation** before starting the processing (see "Adequate, relevant and limited part of the Data minimization" in Part II, section "Principles").

Data minimization might be complex in the case of deep learning, where discrimination by features might be impossible. There is an efficient way to regulate the amount of data gathered and increase it only if it seems necessary: the learning curve. You should start by gathering and using a restricted amount of training data, and then monitor the model's accuracy as it is fed with new data.

2.2.2 Checking legitimate dataset usage

Datasets can be obtained in different ways. Firstly, the developer might opt for acceding to a database that has already been built by someone else. If this is the case, you should be particularly careful, since there are a lot of legal issues that relate to the acquisition of access to database (see "How to access to a database" in Part II, section "Main Tools and Actions").[199]

Secondly, the most common alternative to this consists of building a database. Quite obviously, in this case you have to ensure that you comply with all legal requirements

[198] Article 24.

[199] Yeong Zee Kin, Legal Issues in AI Deployment, At: https://lawgazette.com.sg/feature/legal-issues-in-ai-deployment/ Accessed 15 May 2020

imposed by the GDPR to create a database (see "Creating a database" in Part II, section "Main Tools and Actions").

Thirdly, you might choose an alternative path. You can **mix different datasets so as to create a huge training dataset and another one for validation purposes.** This could bring some issues, such as for example the possibility that the combination of these personal data provides some additional information about the data subjects. For instance, it could allow you to identify data subjects, something that was previously not possible. That could involve deanonymizing anonymized data and creating new personal information that was not contained in the original data set, a circumstance that would bring dramatic ethical and legal issues. For instance, "if data subjects gave informed consent for the processing of personal information in the original data sets for particular purposes, they did not necessarily by extension also give permission for the merging of data sets and for data mining that reveals new information. New information produced in this way may also be based on probabilities or conjectures, and therefore be false, or contain biases in the portrayal of persons."[200] Therefore, you should try to avoid such consequences by ensuring that merging datasets do not work against data subject's rights and interests.

Finally, if you use several datasets that pursue different purposes, you should implement adequate measures to separate the different processing activities. Otherwise you could easily use data collected for on purpose to different activities. This might bring issues related to the purpose limitation principle (see "Purpose limitation principle" within Part II section "Principles" of these Guidelines).

2.2.3 Selecting appropriate legal basis

You should decide the legal basis that you will use for processing before starting it, document their decision privacy notice (along with the purposes) and include the reasons why you have made such choices (see "Accountability" in Part II, section "Principles").

You should select the **legal basis that most closely reflects the true nature of your relationship with the individual and the purpose of the processing.** This decision is key, since changing the legal basis for processing is not possible if there are not solid reasons that justify it (see "Purpose limitation" in Part II, section "Principles").

In the case of an AI tool involving patients' data, developers usually feel tempted to use consent as the legal grounds for processing. This could make a sense if you are re-using data that was already gathered for another purpose and consent was the basis that allowed the primary use of the data. Indeed, the GDPR allows the reuse of data for scientific purposes and article 5.1 (b) states that further processing for scientific research purposes shall not be considered to be incompatible with the initial purposes ('purpose limitation'). Thus, in principle, you could re-use those data on the basis of the original consent. However, you must keep in mind that, according to article 9.4 of the

[200] SHERPA, Guidelines for the Ethical Development of AI and Big Data Systems: An Ethics by Design approach, 2020, p 38. At: https://www.project-sherpa.eu/wp-content/uploads/2019/12/development-final.pdf Accessed 15 May 2020

GDPR, "Member States may maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health." Thus, it might well happen that your relevant national regulation introduces exceptions or specific conditions to the re-use of personal data. In any case, you should always remember that your information duties remain. You should provide the data subject, prior to any further processing of their data, with information on that other purpose and any further relevant information as referred to in paragraph 2 of article 13 GDPR.

The discussion about the re-use of data

At the present moment, there is a lively discussion about the re-use of data for research purposes. According to article 5.1 (b) of the GDPR, further processing for scientific purposes shall not be considered incompatible with the initial purposes. Thus, unless your national regulation states different, you can re-use the data available for research purposes, since these are compatible with the original purpose they were collected for.

However, the EDPS argued that, "in order to ensure respect for the rights of the data subject, the compatibility test under Article 6(4) should still be considered prior to the reuse of data for the purposes of scientific research, particularly where the data was originally collected for very different purposes or outside the area of scientific research. Indeed, according to one analysis from a medical research perspective, applying this test should be straightforward"[201]. According to this interpretation, you should only re-use persona data if the circumstances of article 6.4 apply.

This interpretation somehow contradicts the interpretation of this issue by the EDPB, which stated that Article 5(1)(b) GDPR provides that where data is further processed for scientific purposes, "these shall a priori not be considered as incompatible with the initial purpose, provided that it occurs in accordance with the provisions of Article 89, which foresees specific adequate safeguards and derogations in these cases. Where that is the case, the controller could be able, under certain conditions, to further process the data without the need for a new legal basis. These conditions, due to their horizontal and complex nature, will require specific attention and guidance from the EDPB in the future. For the time being, the presumption of compatibility, subject to the conditions set forth in Article 89, should not be excluded, in all circumstances, for the secondary use of clinical trial data outside the clinical trial protocol for other scientific purposes"[202].

Therefore, the situation remains unclear at this moment, even though we consider that the interpretation by the EDPB makes more sense and will probably prevail in the future.

[201] EDPS, A Preliminary Opinion on data protection and scientific research, 6 January 2020, p. 23.

[202] EDPB, Opinion 3/2019 concerning the Questions and Answers on the interplay between the Clinical Trials Regulation (CTR) and the General Data Protection regulation (GDPR) (art. 70.1.b)) Adopted on 23 January 2019, p. 8.

If you can collect new data for your research, we recommend that you avoid consent as the legal basis, especially if data are collected in a situation where patients are in need of urgent health care, as in the case, for example, that they are suffering symptoms associated with COVID. In the context of clinical trials, the EDPB[203] has stated that "it must be kept in mind that even though conditions for an informed consent under the CTR are gathered, a clear situation of imbalance of powers between the participant and the sponsor/investigator will imply that the consent is not "freely given" in the meaning of the GDPR. As a matter of example, the EDPB considers that this will be the case when a participant is not in good health conditions, when participants belong to an economically or socially disadvantaged group or in any situation of institutional or hierarchical dependency. Therefore, and as explained in the Guidelines on consent of the Working Party 29, consent will not be the appropriate legal basis in most cases, and other legal bases than consent must be relied upon (see below alternative legal bases). Consequently, the EDPB considers that data controllers should conduct a particularly thorough assessment of the circumstances of the clinical trial before relying on individuals' consent as a legal basis for the processing of personal data for the purposes of the research activities of that trial."

From our point of view, this opinion might be extended to other scenarios where the power balance is biased. However, it might happen that the corresponding ethics committee does not share our criterion. Please be aware of such circumstances and try to avoid possible inconveniences in advance by consulting the committee and/or your DPO and the supervising authorities if need be.

3 Data preparation

3.1 Description

"The data preparation phase covers all activities to construct the final data set or the data that will be fed into the modeling tool(s) from the initial raw data. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools. The five steps in data preparation are the selection of data, the cleansing of data, the construction of data, the integration of data, and the formatting of data."[204]

This stage includes all activities needed to construct the final dataset that is fed into the model, from initial raw data. It involves the following five tasks, not necessarily performed sequentially:

---

203 OPINION 3/2019 CONCERNING THE QUESTIONS AND ANSWERS ON THE INTERPLAY BETWEEN THE CLINICAL TRIALS REGULATION (CTR) AND THE GENERAL DATA PROTECTION REGULATION (GDPR), AT: HTTPS://EDPB.EUROPA.EU/OUR-WORK-TOOLS/OUR-DOCUMENTS/DICTAMEN-ART-70/OPINION-32019-CONCERNING-QUESTIONS-AND-ANSWERS_EN

204 Colin Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, p. 16.

1. Select data. Decide on the data to be used for analysis, based on relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types.

2. Clean data. Raise data quality to a required level, for example by selecting clean subsets of the data, insertion of defaults, and estimation of missing data by modeling.

3. Construct data. The construction of new data through the production of derived attributes, new records, or transformed values for existing attributes.

4. Integrate data. Combine data from multiple tables or records to create new records or values.

5. Format data. Make syntactic modifications to data that might be required by the modeling tool.

3.2 Main actions that need to be addressed

3.2.1 Introducing safeguards foreseen in Art. 89 of GDPR

Since you are using data for scientific purposes, you must prepare them according to the safeguards foreseen by the GDPR in its article 89. If the purposes of your research can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, i.e., via pseudonymization, those purposes should be fulfilled in that manner. If this is not possible, you must introduce safeguards ensuring that technical and organizational measures that enable an adequate implementation of the principle of data minimization. Please consider the concrete rules established by your national regulation regarding safeguards. Consult with your DPO.

3.2.2 Ensuring accuracy of processing of personal data

According to the GDPR, data must be accurate (see "Accuracy" in Part II, section "Principles").

This means that data are correct and up to date, but also refers to the accuracy of the analytics performed. The EDPB has highlighted the importance of the accuracy of the profiling or the (not exclusively) automated decision-making process at all stages (from the collection of the data to the application of the profile to the individual).[205]

[205] *Guidelines* on *Automated* individual *decision-making and Profiling* for the purposes of Regulation 2016/679 (wp251rev.01). 22/08/*2018, p.* 13; Ducato, Rossana, Private Ordering of Online Platforms in Smart Urban Mobility The Case of Uber's Rating System, CRIDES Working Paper Series no. 3/20202 February 2020 Updated on 26 July 2020, p. 20-21, at: https://poseidon01.ssrn.com/delivery.php?ID=247104118003073117118086021112071111102048023015008020118084071112086000027097102088036101006014057116105116119119026079007006118044033055000114023106007076115096073024007094081002078064098028091093003078095099082108113086098120001079015123027083125024&EXT=pdf&INDEX=TRUE

Controllers are responsible of ensuring accuracy. Therefore, once you have finished with the collection of data, you should implement adequate tools to guarantee the accuracy of the data. This typically involves that you have to make some fundamental decisions on the technical and organizational measures that will render this principle applicable (see Related technical and organizational measures subsection in the Accuracy section in Principles chapter). Since most of data come from patients and most of them are quantitative, you can assume that they are accurate. In any case, accuracy requires an adequate implementation of measures devoted to facilitate the data subjects' right to rectification (see "Right to rectification" in Part II, section "Data Subject Rights".

3.2.3 Focusing on profiling issues

In the case of a database that will serve to train or validate an AI tool, there is a particularly relevant obligation to inform the data subjects that **their data might cause automated decision-making or profiling on them, unless you can guarantee that the tool will in no way produce these consequences.** Even though automatic decision-making can hardly happen in the context of research, developers should keep an open eye on this issue. Profiling, on the other hand, might bring some problems to AI development.

According to Article 22(3), automated decisions that involve special categories of personal data, such as the health data that you are using, are permitted only if the data subject has consented, or if they are conducted on a legal basis. This exception applies not only when the observed data fit into this category, but **also if the alignment of different types of personal data can reveal sensitive information about individuals or if inferred data enter into that category.**

Some additional actions that might be extremely useful to avoid profiling if it is not needed are:

- Consider the system requirements necessary to support a meaningful human review **from the design phase.** Particularly, the interpretability requirements and effective user-interface design to support human reviews and interventions;
- Design and deliver appropriate training and support for human reviewers; and
- Give staff the appropriate authority, incentives and support to address or escalate individuals' concerns and, if necessary, override the AI tool's decision.

If you proceed with profiling or automated decisions, you must inform the data subjects about your decision and provide all necessary information according to the GDPR and national regulation, if applicable.

3.2.4 Selecting non-biased data

Bias is one of the main issues involved in AI development, an issue that contravenes the fairness principle (see "Lawfulness, fairness and transparency principle" within Part II section "Principles" of these Guidelines). Bias might be caused by a lot of different issues. When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. Sometimes, it might happen that datasets are biased due to malicious actions. Feeding malicious data into an AI tool may change its behavior,

particularly with self-learning systems.[206] Therefore, issues related to the composition of the databases used for training raise crucial ethical and legal issues, not only issues of efficiency or of a technical nature.

You need to address these issues prior to training the algorithm. Identifiable and discriminatory bias should be removed in the dataset building phase where possible. In the case of COVID, distinctions could be made between patients depending on their age, genre, or ethnic group, for instance. You must ensure that the algorithm takes this factor into consideration when you select the data. This means that **the teams in charge of selecting the data to be integrated in the datasets should be composed of people that ensure the diversity that the AI development is expected to show.** Finally, always keep in mind that, if your data are mainly related to a concrete group, for example the Caucasian population more than forty years old, you shall declare that the algorithm has been trained on this basis and, thus, it might not work as well in other population groups.

4 Modeling (training)

4.1 Description

"In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, several techniques exist for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models."[207]

This phase involves several key tasks. Overall, you must:

- Select the modeling technique that will be used. Depending on the type of technique, consequences such as data inference, obscurity or biases are more or less likely to happen.
- Make a decision on the training tool to be used. This enables the developer to measure how well the model can predict history before using it to predict the future. Training always involves running empirical testing with data. Sometimes, developers test the model with data that are different from those used to generate it. Therefore, at this stage one might talk about different types of datasets. Sometimes identifying the individuals that the training data relates to might be difficult. This creates issues for fulfilling individuals' rights that should be addressed appropriately.

---

[206] High-Level Expert Group on AI, Ethics guidelines for trustworthy AI, 2019, p. 17. At: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai Accessed 15 May 2020

[207] Colin Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, p. 17.

4.2 Main actions that need to be addressed

4.2.1 Implementing data minimization principle

According to the data minimization principle, you must proceed to reduce the amount of data and/or the range of information about the data subject they provide as soon as possible. Consequently, you have to purge the data used during the training phase of all information not strictly necessary for training of the model. (see "Temporal aspect subsection in Data minimization" in Part II section "Principles"). There are multiple strategies to ensure data minimization at the training stage. Of course, you should start by erasing all personal data related to the X-ray that you use, but this would only be a first step to follow the minimization principle. Stronger measures should be carefully implemented for this purpose. Techniques are continuously evolving. However, some of the most common are[208] (see also "Integrity and confidentiality" in Part II, section "Principles "):

-Analysis of the conditions that the data must fulfil in order to be considered of high quality and with a great predictive capacity for the specific application.

- Critical analysis of the extent of the data typology used in each stage of the AI tool.

- Deletion of unstructured data and unnecessary information collected during the pre-processing of the information.

-Identification and suppression of those categories of data that do not have a significant influence on learning or on the outcome of the inference.

- Suppression of irrelevant conclusions associated with personal information during the training process, for example, in the case of unsupervised training.

- Use of verification techniques that require less data, such as cross-validation.

- Analysis and configuration of algorithmic hyperparameters that could influence the amount or extent of data processed in order to minimize them.

- Use of federated rather than centralized learning models.

- Application of differential privacy strategies.

- Training with encrypted data using homomorphic techniques.

- Data aggregation.

- Anonymization and pseudonymization, not only in data communication, but also in the training data, possible personal data contained in the model and in the processing of inference.

[208] AEPD, Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción, 2020, p.40. At: https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf Accessed 15 May 2020.

4.2.2 Detecting and erasing biases

Even though the mechanisms against biases are conveniently adopted in previous stages (see previous section about training), it is still necessary to ensure that the results of the training phase minimize biases. This can be difficult, since some types of bias and discrimination are often particularly hard to detect. The team members who are curating the input data are sometimes unaware of them, and the users who are their subjects are not necessarily cognizant of them either. Thus, the monitoring systems implemented by the AI developer in the validation stage are extremely important factors to avoid biases.

There are a lot of technical tools that might serve well to detect biases, such as the Algorithmic Impact Assessment.[209] You must consider their effective implementation.[210] However, as the literature shows[211], it might happen that an algorithm cannot be totally purged of all different types of biases. You should however try to at least be aware of their existence and the implications that this might bring (see "Lawfulness, fairness and transparency" and "Accuracy" in Part II, section on "Principles").

4.2.3 Exercising data subjects' rights

Sometimes, developers complete the available data through inference. For instance, if you do not have the concrete data corresponding to the arterial pressure of a patient, you might use another algorithm to infer it from the rest of the data. However, this does not mean that these data can be considered as entirely pseudonymized or anonymized. This is particularly true in the case of genomic data, since their anonymization is almost impossible. Thus, they continue to be personal data. Furthermore, inferred data must also be considered personal data. Therefore, data subjects have some fundamental rights on these data that you must respect.

Indeed, you must facilitate all data subjects' right during the whole life cycle. In this specific stage, right to access, rectification and erasure are particularly sensitive and include certain characteristics that controllers need to be aware of. However, in the case of research for scientific purposes such as the one you are developing, the GDPR includes some safeguards and derogations relating to processing (art. 89). You must be aware of the concrete regulation in your Member state. According to the GDPR, Union or Member State law may provide for derogations from the main rights included in articles 15 and ff. in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes.

---

[209] Reisman, D., Crawford, K., Whittaker, M., Algorithmic impact assessments: A practical framework for public agency accountability, 2018, at: https://ainowinstitute.org/aiareport2018.pdf Accessed 15 May 2020

[210] https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf Accessed 15 May 2020

[211] Chouldechova. Alexandra, Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments, Big Data. Volume: 5 Issue 2: June 1, 2017. 153-163.http://doi.org/10.1089/big.2016.0047

-Right of access (see "Right to access" in Part II, section "Data Subject Rights")

In principle, you shall respond to data subjects' requests to gain access to their personal data, assuming they have taken reasonable measures to verify the identity of the data subject, and no other exceptions apply. However, you do not have to collect or maintain additional personal data to enable identification of data subjects in training data for the sole purposes of complying with the regulation. If you cannot identify a data subject in the training data and the data subject cannot provide additional information that would enable their identification, they are not obliged to fulfil a request that is not possible to satisfy.

-Right to rectification (see "Right to rectification" in Part II, section "Data Subject Rights")

In the case of the right to rectification, you must guarantee the right of rectification of the data, especially those generated by the inferences and profiles drawn up by an AI tool. Even though the purpose of training data is to train models based on general patterns in large datasets and thus individual inaccuracies are less likely to have any direct effect on a data subject, the right to rectification cannot be limited. As a maximum, you could ask for a longer period (two extra months) to proceed with the rectification if the technical procedure is particularly complex (art. 11(3)).

-Right to erasure (see "Right to erasure" in Part II, section "Data Subject Rights")

Data subjects hold a right to request the deletion of their personal data. However, this right might be limited if some concrete circumstances apply. According to the ICO, "organizations may also receive requests for erasure of training data. Organizations must respond to requests for erasure, unless a relevant exemption applies and provided the data subject has appropriate grounds. For example, if the training data is no longer needed because the ML model has already been trained, the organization must fulfil the request. However, in some cases, where the development of the system is ongoing, it may still be necessary to retain training data for the purposes of re-training, refining and evaluating an AI tool. In this case, the organization should take a case-by-case approach to determining whether it can fulfil requests. Complying with a request to delete training data would not entail erasing any ML models based on such data, unless the models themselves contain that data or can be used to infer it."[212]

5 Evaluation (validation)

[212] ICO, Enabling access, erasure, and rectification rights in AI tools, At: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/ Accessed 15 May 2020

5.1 Description

"Before proceeding to final deployment of the model built by the data analyst, it is important to more thoroughly evaluate the model and review the model's construction to be certain it properly achieves the business objectives. Here it is critical to determine if some important business issue has not been sufficiently considered. At the end of this phase, the project leader then should decide exactly how to use the data mining results. The key steps here are the evaluation of results, the process review, and the determination of next steps."[213]

This phase involves several tasks that raise important data protection issues. Overall, you must:

- Evaluate the results of the model, for instance, whether it is accurate or not. To this purpose, the AI developer might test it in the real world.
- Review the process. You shall review the data mining engagement to determine if there is any important factor or task that has somehow been overlooked. This includes quality assurance issues.

5.2 Main actions that need to be addressed

5.2.1 Processes of dynamic validation

The validation of the processing including an AI component must be done in conditions that reflect the actual environment in which the processing is intended to be deployed. Thus, if you know in advance where the AI tool will be used, you should adapt the validation process to that environment. For instance, if the tool will be deployed in Italy, you should validate it with data obtained from the Italian population, or, if not possible, a similar population. Otherwise, the results might be utterly incorrect. In any case, you should advise about the conditions of the validation to any possible user.

Moreover, the validation process requires periodic review if conditions change or if there is a suspicion that the solution itself may be altered. For instance, if the algorithm is being fed with data from elderly people, you should assess whether or not this changes its accuracy in a young population. You must make sure that validation reflects the conditions in which the algorithm has been validated accurately.

In order to reach this aim, validation should include all components of an AI tool, including data, pre-trained models, environments and the behavior of the system as a whole. Validation should also be performed as soon as possible. Overall, it must be ensured that the outputs or actions are consistent with the results of the preceding processes, comparing them to the previously defined policies to ensure that they are not violated.[214] Validation sometimes needs gathering new personal data. In other cases,

---

[213] Colin Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, p. 17

[214] High-Level Expert Group on AI, Ethics guidelines for trustworthy AI, 2019, p. 22. At: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai v

controllers use data for purposes others than the original ones. In all these cases, controllers should ensure compliance with the GDPR (see "Purpose limitation" section within "Principles" and "Data protection and scientific research" within "Main Concepts" in Part II of these Guidelines).

5.2.2 Deleting unnecessary datasets

Quite often, the validation and training processes are somehow linked. If the validation recommends improvements in the model, training should be performed again. In principle, once the AI development has finally been achieved, the training stage of the AI tool is completed. At that moment, you should implement the removal of the dataset used for this purpose, unless there is a lawful need to maintain it for the purpose of refining or evaluating the system, or for other purposes compatible with those for which they were collected in accordance with the conditions of Article 6(4) of the GDPR. However, you should always consider that deleting the personal data can work against the need to update the accuracy of tools based on the real-time self-learning of algorithms: if a mistake is found you will probably need to recall the data previously used in the training stage. In the event that data subjects request its deletion, you shall have to adopt a case-by-case approach taking into account any limitations to this right provided by the Regulation (see Art. 17(3)).[215]

5.2.3 Performing external audit of data processing

Since the risks of the system you are developing are high, **an audit of the system by an independent third party must be considered**. A variety of different audits can be used. These might be internal or external, they might cover the final product only, or be performed with less evolved prototypes. They could be considered a form of monitoring or a transparency tool.

In terms of legal accuracy, AI tools must be audited to see whether they process personal data in accordance with provisions of GDPR considering a wide range of issues that might be related with that processing. The High-Level Expert Group on AI stated that "testing processes should be designed and performed by as diverse group of people as possible. Multiple metrics should be developed to cover the categories that are being tested for different perspectives. Adversarial testing by trusted and diverse "red teams" deliberately attempting to "break" the system to find vulnerabilities, and "bug bounties" that incentivize outsiders to detect and responsibly report system errors and weaknesses, can be considered."[216] However, there are good reasons to be skeptical about the capability of an auditor to check the functioning of a machine learning system.

This is why it is sensible to focus on the items included by the AEPD in its recommended checklist: it would be more straightforward to focus on the measures

---

[215] AEPD, Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción, 2020, p.26. At: https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf

[216] High-Level Expert Group on AI, Ethics guidelines for trustworthy AI, 2019, p. 22. At: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai Accessed 15 May 2020

implemented to avoid bias, obscurity, hidden profiling, etc., and the adequate use of tools such as the DPIA, which can be performed multiple times, than trying to have an in depth understanding of the functioning of a complex algorithm. Implementing adequate data protection policies form the first stages of the lifecycle of the tool is the best way to avoid data protection issues.

5.2.4 Ensuring compliance with legal framework for medical devices

Before deploying your device, you should ensure that you have adequately followed the regulation regarding the development of medical devices. Please, make sure that this is the case. A Clinical Evaluation and a Performance Evaluation should also be developed. The Guidance on Clinical Evaluation (MDR) / Performance Evaluation (IVDR) of Medical Device Software (https://ec.europa.eu/docsroom/documents/40323 could be an excellent tool to this purpose.)

5.2.5 Informing health care workers participating in development about possible issues

It is often the case that AI mechanisms are validated by comparing their performance with that of human elements, in this case, health care professionals. This can surreptitiously lead to their participation inducing an evaluation of their own professional ability. If we compare the success rate of some professionals with others, some of them may feel that they are being inadvertently tested. It is very important to try to avoid this effect. If it is going to occur, participants should be warned and accept this.

6 Deployment

6.1 Description

"Deployment is the process of getting an IT system to be operational in its environment, including installation, configuration, running, testing, and making necessary changes. Deployment is usually not done by the developers of a system but by the IT team of the customer. Nevertheless, even if this is the case, developers will have a responsibility to supply the customer with sufficient information for successful deployment of the model. This will normally include a (generic) deployment plan, with necessary steps for successful deployment and how to perform them, and a (generic) monitoring and maintenance plan for maintenance of the system, and for monitoring the deployment and correct usage of data mining results."[217]

[217] SHERPA, Guidelines for the Ethical Development of AI and Big Data Systems: An Ethics by Design approach, 2020, p 13. At: https://www.project-sherpa.eu/wp-content/uploads/2019/12/development-final.pdf Accessed 15 May 2020

6.2 Main actions that need to be addressed

6.2.1 General remarks

Once you have created your algorithm, you face an important issue. It might happen that it incorporates personal data, openly or in a hidden way. You must perform a formal evaluation assessing which personal data from the data subjects could be identifiable. This can be complicated at times. For example, some AI tools, such as Vector Support Machines (VSM) could contain examples of the training data by design within the logic of the model. In other cases, patterns may be found in the model that identifies a unique individual. In all of these cases, unauthorized parties may be able to recover elements of the training data, or infer who was in it, by analyzing the way the model behaves. If you know or suspect that the AI tool contains personal data (see "Purchasing or promoting access to a database" section in "Main Tools and Actions", Part II of these Guidelines), you should:

- Delete them or, on the contrary, to justify the impossibility of doing so, completely or partly because of the degradation it would mean for the model (see "Storage limitation section in "Principles" within Part II).

- Determine the legal basis for carrying out the communication of personal data to third parties, especially if special categories of data are involved (see "Lawfulness" subsection in "Lawfulness, fairness and transparency" within "Principles" in Part II).

- Inform the data subjects of the processing above.

- Demonstrate that the data protection by design and by default policies have been implemented (especially data minimization) (see "Data protection by design and by default" within Part II section "Main concepts" of these Guidelines).

- Conduct a Data Protection Impact Assessment (DPIA) (see "DPIA" within Part II section "Main actions and tools" of these Guidelines)

Finally, you must take regular action to proactively evaluate the likelihood of the possibility of personal data being inferred from models in light of the state-of-the-art technology, so that the risk of accidental disclosure is minimized. If these actions reveal a substantial possibility of data disclosure, necessary measures to avoid it should be implemented (see "Integrity and confidentiality" section in "Principles" within Part II of these Guidelines).

6.2.2 Updating information

If the algorithm is implemented by a third party, you must communicate the results of the validation and monitoring system employed at the development stages and offer your collaboration to continue monitoring the validation of the results. It would also be advisable to establish this kind of coordination with third parties from whom you acquire databases or any other relevant component in the life cycle of the system. If this

involves data processing by a third party, you must ensure that access is provided within a legal basis.

It is necessary to offer real time information to the end user about the values of accuracy and/or quality of the inferred information at each stage (see "Accuracy" section in "Principles", Part II of these Guidelines). When the inferred information does not reach minimum quality thresholds, you must highlight that this information has no value. This requirement often implies that you shall provide detailed information about the training and validation stages. Information about the datasets used for those purposes is particularly important. Otherwise, the use of the solution might bring disappointing results to the end users, who are left speculating on the cause.

Second scenario: AI for Crime Prediction and Prevention

Johann Čas (ITA/OEAW)

This part of The Guidelines has been reviewed and validated by Marko Sijan, Senior Advisor Specialist, (HR DPA)

Introduction and preliminary remarks

The use of advanced ICTs plays – as an essential technology for all economic, governmental or societal activities – an increasingly important role in predicting, preventing, investigating, and prosecution of criminal or terroristic activities. Accordingly, research to develop and improve technical capabilities of law enforcement agencies (LEAs) forms a priority area of past, current and future EC funding programs. Advanced and emerging ICTS possess unprecedented powers of surveillance and analysis of large and diverse datasets, particularly in connection with AI[218]

[218] AI is a (too) frequently used term lacking a unique definition. Here we refer to the broad definition of AI, developed by the High-Level Expert Group on AI:
"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI tools can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions.

technologies. The research in such technologies, as well as the implementation of advanced ICTs in the context of security, raise serious concerns of ethics and legal compliance. EU funded security research programs demand explicitly full compliance with the provisions of the Charter of Fundamental Rights of the European Union,[219] the consideration of privacy by design, data protection by design, privacy by default and data protection by default, [220] and, in addition to the Ethics Self-Assessment Table[221] also to fill in a Societal Impact Table. "A 'Societal Impact Table' is a specific feature of this work program part. This table emphasizes on societal aspects of security research. It checks whether the proposed security research meets the needs of and benefits society and does not negatively impact society. Applicants must fill in the 'Societal Impact Table' as part of the submission process." [222] Similar procedures should also be implemented on the level of designing the work programs. Additional safeguards should be foreseen that programs do not contain calls that are difficult or impossible to fulfil without raising severe ethics issues or causing unproportionate infringements of human rights. This could be realized by a mandatory involvement of civil society representatives and ethics and legal expertise among the expert groups drafting EU funded research programs.

These precautions are essential to bringing security research in line with principles like human rights and democracy; nevertheless, concerns remain that they may increase the legitimacy of security research projects without guaranteeing ethical and legal compliance in practice.[223] The use of AI in the context of crime prediction or prevention poses severe threats to civil liberties. A simple trade-off between security and freedom is not appropriate or sufficient. The complex relationship should be treated as a kind of hostile symbiosis,[224] implying that both are necessary for the survival of the other.

To take these concerns into due consideration, this scenario also incorporates information from existing H2020 Security research calls, particularly of the H2020-

---

As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)." https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

[219] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2010:083:0389:0403:en:PDF

[220] For details see EDPB. (2019). Guidelines 4/2019 on Article 25 Data Protection by Design and by Default Version 2.0. Adopted on 20 October 2020. <https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf>

[221] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-self-assess_en.pdf

[222] See p.5 https://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-security_en.pdf

[223] Leese, M., Lidén, K. und Nikolova, B., 2019, Putting critique to work: Ethics in EU security research, Security Dialogue 50(1), 59-76 < https://journals.sagepub.com/doi/abs/10.1177/0967010618809554 >.

[224] Wittes, B. (2011). Against a Crude Balance: Platform Security and the Hostile Symbiosis Between Liberty and Security. Project on Law and Security, Harvard Law School and Brookings, <https://www.brookings.edu/wp-content/uploads/2016/06/0921_platform_security_wittes.pdf>

SEC-2016-2017 call and currently running or recently concluded projects. MAGNETO[225] (Multimedia Analysis and Correlation Engine for Organised Crime Prevention and Investigation), CONNEXIONs[226] (InterCONnected NEXt-Generation Immersive IoT Platform of Crime and Terrorism DetectiON, PredictiON, InvestigatiON, and PreventiON Services) or RED-Alert[227] (Real-time Early Detection and Alert System for Online Terrorist Content based on Natural Language Processing, Social Network Analysis, Artificial Intelligence and Complex Event Processing) are examples of projects of relevance for this case study. They are financed by the 2016-2017 Technologies for prevention, investigation, and mitigation in the context of the fight against crime and terrorism call.[228] The original plan to take one of these projects as concrete bases for this scenario was abandoned as most, or almost all deliverables of the mentioned projects are in accordance with H2020 regulations[229] classified and not publicly accessible. Whereas the classification of specific results of security research projects may be necessary and understandable, it certainly also limits the possibility of public scrutiny and debates of these technologies, which should be mandatory in view of the potential infringements of human rights and European values.

The complexity of this use case is further increased by the fact that different regulations apply to the research and development phase on the one hand and to the implementation and use phase, on the other hand. Research activities are subject to the GDPR; future applications of the research results are subject to the so-called *Data Protection Law Enforcement Directive* (Directive 2016/680),[230] allowing for specific implementation and legislation in individual member states.

The development of AI for security objectives demands particularly careful and strict consideration and compliance with ethics requirements in general, i.e. the already mentioned Horizon 2020 Programme Guidance – How to complete your ethics self-assessment, of respective key documents related to AI, e.g. the High-Level Expert Group on AI: 'Ethics guidelines for trustworthy AI'[231] and the EU Commission White Paper on Artificial Intelligence – A European approach to excellence and trust,[232] and of additional, security specific considerations and documents, as addressed in the Societal Impact Table, the EGE Opinion n°28 – Ethics of Security and Surveillance

[225] http://www.magneto-h2020.eu/
[226] https://www.connexions-project.eu/
[227] https://redalertproject.eu/
[228] https://cordis.europa.eu/programme/id/H2020_SEC-12-FCT-2016-2017
[229] http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/secur/h2020-hi-guide-classif_en.pdf
[230] European Parliament and the Council, 2016, DIRECTIVE (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, Official Journal <http://eur-lex.europa.eu/legal-content/EL/TXT/?uri=OJ:L:2016:119:TOC >.
[231] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[232] https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Technologies[233] or relevant documents published by the EDPS (European Data Protection Supervisor).[234] The Proposal for an Artificial Intelligence Act specifically addresses the use of AI technologies for the purpose of law enforcement and "…lays down a solid risk methodology to define "high-risk" AI tools that pose significant risks to the health and safety or fundamental rights of persons. Those AI tools will have to comply with a set of horizontal mandatory requirements for trustworthy AI and follow conformity assessment procedures before those systems can be placed on the Union market."[235] Annex III lists a number of AI uses for law enforcement as High Risk AI tools for which conformity assessment procedures are mandatory.

The following step by step analysis follows the structure and terminology of the CRISP-DM Model[236] , as outlined in the description below. In order to increase comparability of the approaches and results, this structure is commonly applied to all case studies presented and discussed as part of the MLEs (Mutual Learning Encounters) conducted by the PANELFIT. The adoption of a common structure implies that the individual terms must not be understood literally. Business understanding might for instance, signify to develop a holistic view of the project objectives and on the means and steps to attain them in the case that the planned project does not (primarily) have commercial intentions. It also implies that some of the steps or tasks included in the common framework are not applicable or less relevant for different contexts of the case studies. For instance, the first of the four main tasks comprising the general objective, i.e. the determination of the business objectives, is characterized by little or less freedom of choice if the goals are defined and described in a call to submit research proposals, as it is the case here. This statement should however, not infer that freedoms of choice do not exist at all or should not be considered, but that available options for project applicants are limited in comparison to those available when deciding on the topics of research calls.

During the discussion of the draft version with external experts, we also received recommendations going beyond this specific scenario, e.g. developing curricula for ethics and their mandatory integration into technical studies or offers of data protection and ethics training for engineers. Corresponding training programs should also be offered to police forces (deploying AI) as a general awareness-raising activity.

Step by Step Analysis

[233] European Group on Ethics in Science and New Technologies. (2014). Opinion No. 28: Ethics of security and surveillance technologies (10.2796/22379). Retrieved from Luxembourg: Brussels: https://publications.europa.eu/en/publication-detail/-/publication/6f1b3ce0-2810-4926-b185-54fc3225c969/language-en/format-PDF/source-77404258

[234] https://edps.europa.eu/data-protection/our-work/subjects_en

[235] European Commission. (2021). COM(2021) 206 final. 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., p. 3 <https://ec.europa.eu/newsroom/dae/redirection/document/75788>

[236] Shearer, Colin, The CRISP-DM Model: The New Blueprint for Data Mining, p. 14.

1 Business understanding

1.1 Description

"The initial business understanding phase focuses on understanding the project objectives from a business perspective, converting this knowledge into a data mining problem definition, and then developing a preliminary plan designed to achieve the objectives. In order to understand which data should later be analyzed, and how, it is vital for data mining practitioners to fully understand the business for which they are finding a solution. The business understanding phase involves several key steps, including determining business objectives, assessing the situation, determining the data mining goals, and producing the project plan."[237]

In the context of R&D on crime prediction and prevention technologies conducted in the framework of H2020, the general description and structure of tasks must be adjusted accordingly. This may imply that both the terminology and the concrete contents of the task needs to be interpreted and modified to fit the particular objectives.

The above-mentioned general objectives involve four main tasks:

1. Determine the Project Objectives. This means:
    a. Uncover the primary objectives as well as the related questions the project (envisaged solution) would like to address
    b. Determine the measure of success.
2. Assess the Situation
    a. Identify the resources available to the project, both material and personal.
    b. Identify what data is available to meet the primary goal.
    c. List the assumptions made in the project.
    d. List the project risks, list potential solutions to those risks, create a glossary of project and data processing terms, and construct a cost-benefit analysis for the project.
3. Determine the Data Processing Goals: decide what level of predictive accuracy is expected to consider the project successful.
4. Produce a Project Plan: Describe the intended plan for achieving the data processing goals, including outlining specific steps and a proposed timeline. Provide an assessment of potential risks and an initial assessment of the tools and techniques needed to support the project.

[237] Shearer, Colin, The CRISP-DM Model: The New Blueprint for Data Mining, p. 14.

1.2 Main actions that need to be addressed

1.2.1 Defining project objectives

For our scenario, the general objectives are defined by the respective call. The projects mentioned above pertain to the SEC-12-FCT-2016-2017 call: Technologies for prevention, investigation, and mitigation in the context of the fight against crime and terrorism.[238] The Specific Challenge is described as "Organized crime and terrorist organizations are often at the forefront of technological innovation in planning, executing and concealing their criminal activities and the revenues stemming from them. Law Enforcement agencies (LEAs) are often lagging behind when tackling criminal activities supported by "advanced" technologies".

The scope of this call comprises:

- New knowledge and targeted technologies for fighting both old and new forms of crime and terrorist behaviors supported by advanced technologies;
- Test and demonstration of newly developed technology by LEAs involved in proposals;
- Innovative curricula, training and (joint) exercises to be used to facilitate the EU-wide take-up of these new technologies, in particular in the fields of the following sub-topics:

  1. cyber-crime: virtual/crypto currencies des-anonymization/tracing/ impairing where they support underground markets in the darknet.

  2. detection and neutralization of rogue/suspicious light drone/UAV flying over restricted areas, and involving as beneficiaries, where appropriate, the operators of infrastructure

  3.video analysis in the context of legal investigation

  and a fourth open sub-topic.

The conditions set in this call allow for some, although limited, discretion to design the project. The applicants are free to choose the type of technologies; however, non-technical solutions strategies appear not to be eligible for funding. Even though the range of technologies remains open, the call clearly demands technical solutions, thus excluding approaches to solve the addressed specific security problems without the involvement of potentially highly intrusive technologies. The term advanced technologies at least suggests investigating into developing and using artificial intelligence and machine learning technologies. Limited choice also exists regarding the objective, e.g. on which forms of crime or terrorist behaviors the project targets. Therefore, it is essential to involve end-users, i.e. LEAs (law enforcement agencies), already in the decision-making phase on objectives and the means to achieve them.

The selection of specific technologies, or in a more general context, of specific methods, also influences the range of ethics or legal compliance issues involved by the

[238] https://ec.europa.eu/research/participants/data/ref/h2020/wp/2016_2017/main/h2020-wp1617-security_en.pdf

project. In the case of security research, specifically selected technologies, in our case particular AI or machine learning approaches, may, apart from usual ethics issues like the processing of personal data, raise in addition ethics concerns related to dual-use, the exclusive focus of the research on civil applications or to misuse, requiring to consider related particular regulations accordingly.

1.2.2 Opting for technical solutions with explainability and transparency

Whereas explainability and transparency constitute generic requirements for AI tools, they form mandatory obligations in the case of AI technologies applied to or having consequences for humans (see also the "Lawfulness, fairness and transparency" section within "Principles" in Part II). In the case of AI used for profiling or decision support in a security context, these principles are fundamental. AI tools are prone to bias; explainability and transparency can help detect and remove biases of algorithms created by such methods. Technologies supporting crime prevention, detection and prosecution need to provide provable and attestable results as valid evidence, also in front of courts. Inaccurate findings may have severe consequences for individuals, particularly in the form of false positives or fatal outcomes in the case of false negatives. Therefore, it may be necessary to implement the AI tool as a support for decisions by humans, together with mandatory measures accompanying the employment. Thus, making sure that people in charge do not just make the system's suggestion to their own decision, but understand that they have to justify their decision, when following the suggestion as well as when objecting a suggestion by the system. To enable humans to understand the suggestion of an AI tool, these systems need to be very transparent regarding the factors influencing the outcome of a calculation. In the end, humans need to take responsibility as well as the liability for a decision. Transparency is also essential to ensure sufficient understanding of the model and data used and the results produced, particularly in the case of complaints or need of proof of evidence.

Developers of AI tools used in this context could facilitate the implementation by programming supporting applications for the whole decision process, like having a mandatory field to fill in when a decision is made upon the suggestion by the system before the outcome could be processed further.

1.2.3 Implementing a training program

In our case, "training and (joint) exercises to be used to facilitate the EU-wide take-up of these new technologies" are already included in the call description. Such training exercises must not be restricted to the use of the developed technologies, but start at the very beginning of research activities and in particular, comprise all persons involved in the design of AI technologies (e.g., algorithm designers, developers, programers, coders, data scientists, engineers). This action is one of the essential pieces of advice to be considered from the very first moment of a crime prediction and prevention project. Algorithm designers, who occupy the first link in the algorithmic chain, are likely to be unaware of the ethical and legal implications of their actions. One of the main problems of AI tools devoted to dealing with crime and terrorism is that they often use personal data that are included in large datasets, comprising large fractions of citizens, e.g. users of specific social networks. Whereas the analysis of mass surveillance data by AI tools may be permissible under specific national jurisdictions or transpositions of the *Data*

Protection Law Enforcement Directive (Directive 2016/680), it is still very problematic for several reasons. First, legal compliance may be necessary condition for conformity with ethics principles, but never can be regarded as a sufficient condition. An information document provided by the European commission on "Ethics and data protection"[239] clearly states that "The fact that some data are publicly available does not mean that there are no limits to their use" (see Box 4 on page 13). Second, compliance with national or EU legislation does not necessarily imply legal compliance with fundamental rights. The Data Retention Directive[240] is a prominent related example as it was annulled by Court of Justice of European Union (CJEU) in a ruling of 8 April 2014[241] because the Court considered that the directive 'entails a wide-ranging and particularly serious interference with the fundamental rights to the respect for private life and to the protection of personal data, without that interference being limited to what is strictly necessary'. Third, public opinion and acceptability by citizens must be respected. Large-scale citizen consultations on surveillance technologies revealed that citizens in general accept serious intrusions into their privacy if they are based on concrete and plausible suspicion but reject untargeted mass surveillance measures.[242] Applying data mining to detect criminal or terroristic activities can be compared to finding the needle in the haystack[243]. This also means that the processing will include personal data of data subjects that are not currently or have not been in the past involved in any criminal or terrorist activities. Depending on the targeting of the data analyzed, the data processed may predominantly or almost exclusively concern innocent individuals. Such data processing violates the presumption of innocence, changes the relationship between citizens and state and may have grave societal and individual (in case of false positives) consequences.

You, as an algorithm designer, must therefore be able to understand the implications of your actions, both for individuals and society, and be aware of your responsibilities by learning to show continued attention and vigilance. Following this advice may help you in avoiding or mitigating many ethical and legal issues. In that sense, an optimal training for all subjects involved in the project even before it starts could be one of the most efficient tools to save time and resources in terms of compliance with data

---

239 https://ec.europa.eu/info/sites/default/files/5._h2020_ethics_and_data_protection_0.pdf

240 Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks and amending Directive 2002/58/EC, Official Journal of the European Union <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32006L0024&from=en>

241 https://www.europarl.europa.eu/legislative-train/theme-area-of-justice-and-fundamental-rights/file-data-retention-directive

242 Strauß, S. (2015). *D 6.10–Citizen Summits on Privacy, Security and Surveillance: Synthesis Report*. <http://surprise-project.eu/wp-content/uploads/2015/02/SurPRISE-D6.10-Synthesis-report.pdf>

243 Which also means that searching more data just increases the haystack, not necessarily the number of needles.

protection, ethics, EU and national law or societal acceptability. This also implies the participation of ethical and legal experts both in training activities and in the execution of the project. Adequate measures to ensure confidentiality also deserve particular attention (see "Measures in support of confidentiality" in the "Integrity and confidentiality" section within "Principles" in Part II). Security and confidentiality of processed data, on the one hand, is essential; general knowledge about the types of mined data, persons concerned or algorithms applied, on the other hand, is mandatory to guarantee compliance with human rights and European values. Compliance with the most restrictive member state also supports business objectives, allowing the implementation and use of developed systems without the need for individual adjustments.

1.2.4 Using legal framework applicable for data processing

For security-related R&D projects this step is particularly complex and challenging. For the research project as such GDPR regulations apply; for later implementations the rules and provisions of the *Data Protection Law Enforcement Directive* (Directive 2016/680) must be followed. In addition, possibly diverging legislation of involved (member) states need to be taken into account. Therefore, the developed technologies and systems must at least provide for adjustability and flexibility to cope with different regulations. From a human rights and ethics perspective, compliance with the most restrictive should be incorporated in the created technologies, thus supporting maximum respect for fundamental rights and related values, at the same time, as already mentioned, reducing or eliminating the need for modifications if applied in countries with diverging regulations.

According to article 5(1)(a) of the GDPR, personal data shall be "collected for specific, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes". The concept of legitimacy is not well defined in the GDPR, but the Article 29 Working Party stated that legitimacy involves that data must be processed "in accordance with the law", and "law" should be understood as a broad concept that includes "all forms of written and common law, primary and secondary legislation, municipal decrees, judicial precedents, constitutional principles, fundamental rights, other legal principles, as well as jurisprudence, as such 'law' would be interpreted and taken into account by competent courts".[244]

Therefore, it is a wider concept than lawfulness. It involves compliance with the main values of applicable regulations and the main ethical principles at stake. For instance, some concrete AI tools will need the intervention of an ethics committee. In other cases, guidelines or any other kind of soft regulation might be applicable. You should ensure adequate compliance with this requirement by designing a plan for this preliminary stage of the lifecycle of the tool (see "Legitimacy and lawfulness" in "Lawfulness, fairness and transparency" within "Principles" in Part II). To this purpose, you should be particularly aware of the requirements posed by the applicable regulation at the national level. Developing algorithms related to crime prediction and prevention clearly

[244] Article 29 Working Party (2013) Opinion 03/2013 on purpose limitation Adopted on 2 April 2013, WP203. European Commission, Brussels, p.20. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

requires the involvement of Ethics Committees from an early stage on and according to Art. 35 GDPR to carry out a Data protection impact assessment. As already mentioned, Art. 10 GDPR requires checking whether the processing is authorized by Union or Member State law in the case of processing personal data relating to criminal convictions and offences or related security measures. Make sure that your research plan fits well with all these requirements for both phases, the conduction of the research project and future implementations of the developed systems.

The ethics guidance provided for EU funded research (see footnote 221) constitutes a comprehensive framework for checking ethics compliance, which should be consulted in addition to institutional ethics regulations or codes of conduct, regardless whether your research actually receives funding from the EC. Please be aware that ethics evaluation is not a checklist activity but always also comprises a weighing of potentially conflicting norms. In particular, the application of emerging ICT and having in mind privacy by design in such a sensitive area, requires forward-thinking on both sides, involved researchers and the ethics evaluators.

Even if your project or your research institution is not subject to specific ethic regulations, observation of and compliance with relevant national or EU is essential. As soon as you bring the developed technologies and systems to the market, compliance is essential for both, implementation within the EU and for getting export licences for non-EU commercial exploitation.

1.2.5 Adopting a risk-based thinking approach

The creation of your algorithm will probably involve the use of several special categories of personal data, e.g. political opinions, religious or philosophical beliefs or data concerning a natural person's sex life or sexual orientation in the case of data mining of social networks. Therefore, you must ensure that you implement appropriate measures to minimize the risks to data subjects' rights, interests, and freedoms (see "Integrity and confidentiality principle" within Part II section "Principles" of these Guidelines). To this purpose, you must assess the risks to the rights and freedoms of individuals participating in the research and development process and judge what is appropriate to protect them. In all cases, you need to ensure compliance with data protection requirements.

In the context of crime prediction, prevention, detection or investigation technologies a risk-based approach makes a DPIA (Data Protection Impact Assessment) obligatory as at least one of the three specific conditions of Art. 35(3) GDPR necessarily will apply:

"*3. A data protection impact assessment referred to in paragraph 1 shall in particular be required in the case of*:

(a) a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person;

(b) processing on a large scale of special categories of data referred to in Article 9(1), or of personal data relating to criminal convictions and offences referred to in Article 10; or

(c) a systematic monitoring of a publicly accessible area on a large scale."

The risk-based analysis should also include potential ethics issues related to misuses[245] of the developed technologies and to dual use[246] related export restrictions that may apply to the developed systems.

Consider also that the risks are not limited to data protection and privacy violating impacts of the developed systems. Constitutional rights and other human rights such as the presumption of innocence, equal access to justice, non-discrimination or freedom of expression may also be violated or impaired. Moreover, these effects are not limited to potential suspects, but affect society as a whole. They are exacerbated by a lack of transparency and human controllability of many AI tools.

1.2.6 Preparing the documentation of processing

Whoever processes personal data (including both, controllers and processors) needs to document their activities primarily for the use by qualified/relevant Supervisory Authorities. You must do this through records of processing that are maintained centrally by your organization across all its processing activities, and additional documentation that pertains to individual data processing activities (see Documentation of Processing section in Actions and Tools chapter). This preliminary stage is the perfect moment to set up a systematic way of collecting the necessary documentation, since it will be the time when you can conceive and plan the processing activity.

The development of your AI tool might involve the use of different datasets. The records must ensure the traceability of the processing, the information about possible reuse of data, and the use of data pertaining to different datasets in different or in the same stages of the life cycle.

For systems used for law enforcement purposes, the documentation of processing must also comprise the documentation of access to the system once implemented in order to prevent and to detect possible misuses, e.g. non-authorized access to generated results.

As stated in the Requirements and acceptance tests for the purchase and/or development of the employed software, hardware, and infrastructure (subsection of the Documentation of Processing section), the risk evaluation and the decisions taken *"have to be documented in order to comply with the requirement of data protection by design (of Art. 25 GDPR). Practically, this can take the form of:*

> *Data protection **requirements** specified for the purchase (e.g., a tender) or development of software, hardware and infrastructure,*

[245] See https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/guide_research-misuse_en.pdf

[246] See https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/guide_research-dual-use_en.pdf

Acceptance tests *that verify that the chosen software, systems and infrastructure are fit for purpose and provide adequate protection and safeguards.*

Such documentation should be an integral part of the DPIA."

Finally, you should always be aware that, according to Art. 32(1)(d) of the GDPR, data protection is a process. Therefore, **you should test, assess, and evaluate the effectiveness of technical and organizational measures regularly.** This stage is a perfect moment to build a strategy aimed at facing these challenges.

1.2.7 Checking regulatory framework

The GDPR includes specific rules regarding processing for the purposes of scientific research (see "Data protection and scientific research" section in "Main Concepts" chapter).[247] Your AI tool might be classified as scientific research, irrespective of whether it is created for profit or not. *"Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes"* (Art. 89(2) GDPR). Furthermore, according to article 5 (b) *"further processing of the data gathered, in accordance with Article 89(1), would not be considered to be incompatible with the initial purposes ('purpose limitation'). Some other particular exceptions to the general framework applicable to processing for research purposes (such as storage limitation) should also be considered".*

Possibly you might profit from this favorable framework, depending on the countries where the research is conducted and on the legal form of the involved partners, e.g. whether they are academic or commercial entities. Nevertheless, you must be aware of the concrete (national) regulations that apply to this research (mainly, the safeguards to be implemented). They might include specific requirements, depending on respective national laws.

To be careful also implies that you have to consider both legal and ethical limitations to the planned research. Just because specific (national) regulations allow for the intended data processing does not imply that it is also acceptable or compliant from an ethics perspective. In analogy ethics compliance must not be misused as an escape[248] from regulations.

[247] This specific framework also includes historical research purposes or statistical purposes. However, ICT research is not usually related to these purposes. Therefore, we will not analyze them here.

[248] Wagner, B. (2018). Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping? In E. BayamlioĞLu, I. Baraliuc, L. Janssens, & M. Hildebrandt (Eds.), Being Profiled (pp. 84-89): Amsterdam University Press.

1.2.8 Defining data storage policies

According to Article 5(1)(e) GDPR, personal data should be *"kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed"*. This requisite is twofold. On the one hand, it relates to identification: data should be stored in a form which permits identification of data subjects for no longer than necessary. Consequently, you should implement policies devoted to avoiding identification as soon as it is not necessary for processing. Such policies involve the adoption of adequate measures to ensure that at any moment, only **the minimal degree of identification that is necessary to fulfil the purposes must be used** (see "Temporal aspect" subsection in "Storage limitation principle" within Part II section "Principles" of these Guidelines).

On the other hand, data storage implies that data can only be stored for a **limited period**: the time that is strictly necessary for the purposes for which the data are processed. However, the GDPR permits 'storage for longer periods if the sole purpose is scientific research' (which might be the case for the R&D phase).

The scientific research exception raises the risk that you decide to keep the data longer than strictly needed. You must be aware that even though the GDPR might allow storage for longer periods, **you should have justifiable reasons to opt for such an extended period.** For the developed systems, you must include organizational and technical precautions to be able to comply with different national legal regulations concerning the maximum data storage periods. This could also be an excellent moment to **envisage time limits for (automatic) erasure of different categories of data and to document these decisions** (see "Accountability principle" within Part II section "Principles" of these Guidelines).

1.2.9 Appointing a Data Protection Officer

According to Art. 37(1) GDPR you must appoint a DPO:

"1. The controller and the processor shall designate a data protection officer in any case where:

> *(a) the processing is carried out by a public authority or body, except for courts acting in their judicial capacity;*
>
> *(b) the core activities of the controller or the processor consist of processing operations which, by virtue of their nature, their scope and/or their purposes, require regular and systematic monitoring of data subjects on a large scale; or*
>
> *(c) the core activities of the controller or the processor consist of processing on a large scale of special categories of data pursuant to Article 9 and personal data relating to criminal convictions and offences referred to in Article 10."*

2 Data Understanding

2.1 Description

"The data understanding phase starts with an initial data collection. The analyst then proceeds to increase familiarity with the data, to identify data quality problems, to discover initial insights into the data, or to detect interesting subsets to form hypotheses about hidden information. The data understanding phase involves four steps, including the collection of initial data, the description of data, the exploration of data, and the verification of data quality".[249]

All of these steps are aimed at identifying the data available. At this stage, you need to be aware of the data you will have to work with and start making decisions on how main principles related to data protection will be implemented. You should consult the Ethics and data protection document from 14 November 2018[250] to comply with legal and ethics requirements. In the case of using data from social networks, the information provided in Box 4 Using 'open source' data on page 13 is particularly relevant.

You should also be aware that databases that contains personal data about prosecutions related to criminal convictions and offenses are sensitive, and that you as developer will normally not be able to access them.

2.2 Main actions that need to be addressed

At this stage, a large number of fundamental issues related to the protection of personal data needs to be addressed. Depending on the decisions made, principles such as data minimization, privacy by design or by default, lawfulness, fairness and transparency, etc. will be adequately settled. A communication between ethics and legal experts, on the one hand, and project developers, on the other hand, has to be established to be able to realize the principles of "privacy by design" or "by default".

2.2.1 Making decision on types of data to be processed

According to the GDPR, the *"*controller *shall implement appropriate technical and organizational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons."*[251] (see Data Protection by Design and by Default in Concepts chapter) This demand must be specially kept in mind during this stage, since decisions about the type of data that will be used are often taken at this moment.

[249] Colin Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, p. 15
[250] https://ec.europa.eu/info/sites/info/files/5._h2020_ethics_and_data_protection_0.pdf
[251] Article 25(2).

Thus, make sure whether you really need vast amounts of data. Focused "smart data" might be much more useful than big data. Of course, using smart, well prepared data might involve a huge effort in terms of unification, homogenization, etc., but it will help to implement the principle of data minimization (see "Data minimization principle" within Part II section "Principles" of these Guidelines) in a much more efficient way. To this purpose**, having expertise available to select relevant features is essential.** This step also involves checking the necessity of processing for each category of data; this implies to prove that no, from a data protection and human rights perspective less infringing, alternative measures or methods could be applied to achieve the same result.

Furthermore, you should try to **limit the resolution of the data** to what is minimally necessary for the purposes pursued by the processing. You should also **determine an optimal level of data aggregation** before starting the processing (see "Adequate, relevant and limited part of the Data minimization" section in "Principles" chapter). In the case of AI applied to crime prediction, prevention or investigation, the possible level of data aggregation, i.e. anonymization of data, is undoubtedly limited, at least for later implementations and uses of the developed systems. As a primary objective is to identify (potential) perpetrators, it must at least be possible to (re-)personalize data on potential threats.

Data minimization might be complicated in the case of deep learning, where differentiation by features might be impossible. There is an efficient way to regulate the amount of data gathered and increase it only if it seems necessary: the learning curve. You should start by collecting and using a limited amount of training data, and then monitor the model's accuracy as it is fed with new data.

2.2.2 Checking legitimate dataset usage

Datasets can be obtained in different ways. Firstly, the developer might opt for acquiring or gaining access to a database that has already been built by someone else. If this is the case, you should be particularly careful since there are a lot of legal issues that relate to the acquisition of access to a database (see "Purchasing access to a database" section in "Main Tools and Actions" chapter).[252]

Secondly, the most common alternative to this consists of building a database. Quite obviously, in this case you have to ensure that you comply with all legal requirements imposed by the GDPR to create a database (see "Creating a database" section in "Main Tools and Actions" chapter).

Thirdly, you might choose an alternative path. You can mix licensed data from third parties with your own dataset so as to create a huge training dataset and another one for validation purposes. This could bring some issues, such as the possibility that the combination of different data sets provides some additional information about the data subjects. For instance, it could allow you to identify data subjects, something that was previously not possible, using only one of the datasets. That could involve de-anonymizing anonymized data and creating new personal information that was not contained in the original data set. This situation would entail significant ethical and

[252] Yeong Zee Kin, Legal Issues in AI Deployment, At: https://lawgazette.com.sg/feature/legal-issues-in-ai-deployment/ Accessed 15 May 2020

legal issues. For instance, *"if data subjects gave informed consent for the processing of personal information in the original data sets for particular purposes, they did not necessarily by extension also give permission for the merging of data sets and for data mining that reveals new information. New information produced in this way may also be based on probabilities or conjectures, and therefore be false, or contain biases in the portrayal of persons."*[253] Therefore, you should try to avoid such consequences by ensuring that merging datasets do not work against data subjects' rights and interests.

Finally, if you use several datasets that pursue different purposes, you should implement adequate measures to separate the different processing activities. Otherwise, you could easily use data for a purpose for which it has not been collected. This might bring issues related to the purpose limitation principle (see "Purpose limitation principle" within Part II section "Principles" of these Guidelines).

Be aware that the above-mentioned measures are only sufficient for the research project execution phase. Informed consent will generally be of very limited use in the context of law enforcement activity. The same holds for the creation and use of dummy or synthetic data. The use of synthetic data still may involve issues of potential re-identification as well as the question of whether one can trust such data when training AI algorithms. All these measures may effectively help to mitigate or eliminate ethics or legal issues for the research phase. It is essential to ensure that the datasets needed for real-world implementations also comply with the ethical and legal requirements imposed by EU and national member state regulations; this also holds for the use of police- or government-owned datasets. Be also aware that it might be difficult or even impossible to get access to sufficient large real datasets required for practical training of the AI tool.

2.2.3 Selecting appropriate legal basis for processing

You must decide the legal basis that you will use for processing before starting it, document your decision (along with the purposes) and include the reasons why you have made your choice (see "Accountability principle" within Part II section "Principles" of these Guidelines).

You should select the legal basis that most closely reflects the true nature of your processing of personal data. In case human participants are involved, also the relationship with the participants and the purpose of the processing must be considered. This decision is key, since changing the legal basis for processing is not possible if there are not solid reasons that justify it (see Purpose limitation section in Principles chapter).

In the case of AI tools developed for the purpose of crime prediction, prevention, etc., you must again distinguish between the research phase and later implementations. For the research phase, you may be able to use consent as the legal ground for processing, depending on the concrete involvement of human participants. Examples could be AI tools using biometric identification or the interpretation of video data, requiring the involvement of human participants for testing. Consent also could form a valid legal

[253] SHERPA, Guidelines for the Ethical Development of AI and Big Data Systems: An Ethics by Design approach, 2020, p 38. At: https://www.project-sherpa.eu/wp-content/uploads/2019/12/development-final.pdf Accessed 15 May 2020

ground if you are reusing data that was already gathered for another purpose and consent was the basis that allowed the primary use of the data. The GDPR allows the reuse of data for scientific purposes and article 5.1 (b) states that further processing for scientific research purposes shall not be considered to be incompatible with the initial purposes ('purpose limitation'). Thus, in principle, you could reuse those data on the basis of the original consent. However, you must keep in mind that, according to article 9.4 of the GDPR, *"Member States may maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health."* Thus, it might well happen that your relevant national regulation introduces exceptions or specific conditions to the reuse of personal data. In any case, you should always remember that your information duties remain. You should provide the data subject, prior to any further processing of their data, with information on that other purpose and any further relevant information as referred to in paragraph 2 of Article 13 GDPR.

Please, keep in mind that the above provisions only hold for conducting the research as such. Future uses of the developed systems need to conform to valid legislation of the EU and of member states concerning law enforcement activities. Also, be aware that developing technologies which are not compliant with applicable regulations or with ethics principles or European values would imply a waste of effort and resources.

2.2.4 Reusing of data

At present, there is a lively discussion about the reuse of data for research purposes. According to article 5.1 (b) of the GDPR, further processing for scientific purposes shall not be considered incompatible with the initial purposes. Thus, unless your national regulation states different, you can reuse the data available for research purposes, since these are compatible with the original purpose they were collected for.

However, the EDPS argues that, *"in order to ensure respect for the rights of the data subject, the compatibility test under Article 6(4) should still be considered prior to the reuse of data for the purposes of scientific research, particularly where the data was originally collected for very different purposes or outside the area of scientific research. Indeed, according to one analysis from a medical research perspective, applying this test should be straightforward".*[254] According to this interpretation, you should only reuse persona data if the circumstances of article 6.4 apply. Please check in this context also the applicability of article 10 *"Processing of personal data relating to criminal convictions and offences or related security measures based on Article 6(1) shall be carried out only under the control of official authority or when the processing is authorized by Union or Member State law providing for appropriate safeguards for the rights and freedoms of data subjects."*

3 Data preparation

[254] EDPS, A Preliminary Opinion on data protection and scientific research, 6 January 2020, p. 23.

3.1 Description

"The data preparation phase covers all activities to construct the final data set or the data that will be fed into the modeling tool(s) from the initial raw data. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools. The five steps in data preparation are the selection of data, the cleansing of data, the construction of data, the integration of data, and the formatting of data."[255]

This stage includes all activities needed to construct the final dataset that is fed into the model, from initial raw data. It involves the following five tasks, not necessarily performed sequentially:

1. Select data: Decide on the data to be used for analysis, based on relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types.
2. Clean data: Raise data quality to a required level, for example by selecting clean subsets of the data, insertion of defaults, and estimation of missing data by modeling.
3. Construct data: The construction of new data through the production of derived attributes, new records, or transformed values for existing attributes.
4. Integrate data: Combine data from multiple tables or records to create new records or values.
5. Format data: Make syntactic modifications to data that might be required by the modeling tool.

3.2 Main actions that need to be addressed

3.2.1 Introducing the safeguards foreseen in Article 89 GDPR

Since you are using data for scientific purposes, you must prepare them according to the safeguards foreseen by the GDPR in Article 89. If the purposes of your research can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, i.e., via pseudonymization, those purposes should be fulfilled in that manner. If this is not possible, you must introduce safeguards ensuring that technical and organizational measures enable an adequate implementation of the principle of data minimization. Please consider the concrete rules established by your national regulation regarding safeguards. Consult with your DPO.

3.2.2 Ensuring accuracy of processing of personal data

According to the GDPR, data must be accurate (see "Accuracy" section in "Principles" chapter). This means that process data are correct and up to date. Controllers are responsible to ensure accuracy. Therefore, once you have finished with the collection of data, you should implement adequate tools to guarantee the accuracy of the data. This typically involves that you have to make some fundamental decisions on the technical

[255] Colin Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, p. 16.

and organizational measures that will render this principle applicable (see "Related technical and organizational measures" subsection in the "Accuracy" section in "Principles" chapter). Since most of the data come from probably quite different sources with no standardised quality requirements and most of them will probably be qualitative in the case of crime prediction, you cannot assume that they are accurate per se. Primarily because these data might be based on individual ratings of different people, while the data subjects might not even know about the fact that this kind of data is stored about them.

In any case, accuracy requires an adequate implementation of measures devoted to facilitate the data subjects' right to rectification (see (see "Right to Rectification" within Part II section "Data subject's rights" of these Guidelines).

Ensure also that they produce results that are as accurate as possible. The types of false positives and false negatives should be defined in advance during the data preparation phase. False results are one of the essential issues having an impact on individuals' fundamental rights.

3.2.3 Focusing on profiling issues

In general, in the case of a database that will serve to train or validate an AI tool, there is a particularly relevant obligation to inform the data subjects that **their data might cause automated decision-making or profiling on them.** Profiling is particularly problematic in AI development, this also holds for AI tools developed for LEAs purposes.

According to Article 22(2)(c), automated decisions that involve special categories of personal data, such as *data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation* (Article 9(1)) are permitted only if the data subject has consented, or if they are conducted on a legal basis. This exception applies not only when the observed data fit into this category, but **also if the alignment of different types of personal data can reveal sensitive information about individuals or if inferred data enter into that category.** In the case of crime prediction and prevention explicit consent from the data subjects will normally only be applicable for voluntarily human participants during the R&D phase. The processing of special categories of personal data, for instance of political opinions or religious beliefs, may belong to the core of data of AI tools applied in the field of terrorism prevention.

Some additional actions that might be extremely useful to avoid automated decision-making if it is not needed are:

- Consider the system requirements necessary to support a meaningful human review **from the design phase.** Particularly, the interpretability requirements and effective user-interface design to support human reviews and interventions;

- Design and deliver appropriate training and support for human reviewers; and
- Give staff the appropriate authority, incentives and support to address or escalate individuals' concerns and, if necessary, override the AI tool's decision.[256]

If you proceed with profiling or automated decisions, you must inform the data subjects about your decision and provide all necessary information according to the GDPR and national regulation, if applicable.

3.2.4 Selecting non-biased data

Bias is one of the main issues involved in AI development, an issue that contravenes the fairness principle (see "Lawfulness, fairness and transparency principle" within Part II section "Principles" of these Guidelines). Bias might be caused by a lot of different issues. When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. Sometimes, it might happen that datasets are biased due to malicious actions. Feeding malicious data into an AI tool may change its behavior, particularly with self-learning systems.[257] Therefore, issues related to the composition of the databases used for training raise crucial ethical and legal issues, not only issues of efficiency or of a technical nature.

You need to address these issues prior to training the algorithm. Identifiable and discriminatory bias should be removed in the dataset building phase where possible. As we have seen in the past, the idea that certain groups of people (Black, Arabs or aliens in general, Muslims…) are convicted more often because they break the law more frequently in most cases is not valid. They are searched more often, discriminated more often by the police, encounter more often excessive violence, arbitrariness or hostility by the police and therefore more often come into problematic situations. This observation would most probably hold for any other subset of the population if treated the same way. Therefore, deducting a higher crime rate in areas where many foreigners live might become a self-fulfilling prophecy.

Another example might be the assumption that an AI tool produces the right results as soon as they match with the results by humans. Often decisions by humans are biased as well, and the AI tool would most probably perpetuate such discriminatory practices instead of producing more objective results.

If the algorithm is biased, it may also increase the number of false positives or false negatives. False positives may have serious adverse effects on concerned individuals, false negatives on society and of course, also on victims of criminal or terroristic activities which could potentially have been avoided.

You must ensure that the algorithm assesses these factors accordingly when you select the data. This means that **the teams in charge of selecting the data to be integrated in**

[256] https://ico.org.uk/about-the-ico/news-and-events/ai-blog-fully-automated-decision-making-ai-systems-the-right-to-human-intervention-and-other-safeguards/

[257] High-Level Expert Group on AI, Ethics guidelines for trustworthy AI, 2019, p. 17. At: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai Accessed 15 May 2020

the datasets should be composed of people that ensure the diversity that the AI tool is expected to show. Finally, always keep in mind that, if your data are mainly related to a concrete group, you shall declare that the algorithm has been trained on this basis and, thus, it might not work as well in other population groups.

4 Modeling (training)

4.1 Description

"In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, several techniques exist for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase may be necessary. Modeling steps include the selection of the modeling technique, the generation of test design, the creation of models, and the assessment of models."[258]

This phase involves several key tasks. Overall, you must

- Select the modeling technique that will be used. Depending on the type of technique, consequences such as data inference, obscurity or biases are more or less likely to happen.
- Make a decision on the training tool to be used. This enables the developer to measure how well the model can predict history before using it to predict the future. In the case of crime prediction, this could be a problem itself. It's not like predicting that someone with love for yoghurt will buy it again. We are talking about human beings and their chances in life. Assuming one would re-offend because they did something illegal in the past almost neglects the fact that we think of citizens as humans with free will and the chance to make a better decision the next time. It is inherently problematic to assume the future will be an extrapolation of the past. Depending on the individual and societal consequences, it might be less of a problem in some cases and unjustifiable in others.

  Training always involves running empirical testing with data. Sometimes, developers test the model with data that are different from those used to generate it. Therefore, at this stage one might talk about different types of datasets.

4.2 Main actions that need to be addressed

4.2.1 Implementing data minimization principle

According to the data minimization principle, you must proceed to reduce the amount of data and/or the range of information about the data subject they provide as soon as possible. Consequently, you have to purge the data used during the training phase of all

[258] Colin Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, p. 17.

information not strictly necessary for the training of the model. (see "Temporal aspect" subsection in "Data minimization" within "Principles" in Part II. There are multiple strategies to ensure data minimization at the training stage. Techniques are continuously evolving. However, some of the most common are[259] (see "Integrity and confidentiality principle" within Part II section "Principles" of these Guidelines):

- Analysis of the conditions that the data must fulfil in order to be considered of high quality and with a great predictive capacity for the specific application.
- Critical analysis of the extent of the data typology used in each stage of the AI solution.
- Deletion of unstructured data and unnecessary information collected during the pre-processing of the information.
- Identification and suppression of those categories of data that do not have a significant influence on learning or on the outcome of the inference.
- Suppression of irrelevant conclusions associated with personal information during the training process, for example, in the case of unsupervised training.
- Use of verification techniques that require less data, such as cross-validation
- Analysis and configuration of algorithmic hyperparameters that could influence the amount or extent of data processed in order to minimize them
- Use of federated rather than centralized learning models
- Application of differential privacy strategies.
- Training with encrypted data using homomorphic techniques.
- Data aggregation.
- Anonymization and pseudonymization, not only in data communication, but also in the training data, possible personal data contained in the model and in the processing of inference.

4.2.2 Detecting and erasing biases

Even though the mechanisms against biases are conveniently adopted in previous stages (see the section about training above), it is still necessary to ensure that the results of the training phase minimize biases. This can be difficult since some types of bias and discrimination are often particularly hard to detect. The team members who are curating the input data are sometimes unaware of them, and the users who are their subjects are not necessarily cognisant of them either. Thus, the monitoring systems implemented by the AI developer in the validation stage are extremely important factors to avoid biases.

There are a lot of technical tools that might serve well to detect biases, such as the Algorithmic Impact Assessment.[260] You must consider their effective

[259] AEPD, Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción, 2020, p.40. At: https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf Accessed 15 May 2020.

[260] Reisman, D., Crawford, K., Whittaker, M., Algorithmic impact assessments: A practical framework for public agency accountability, 2018, at: https://ainowinstitute.org/aiareport2018.pdf Accessed 15 May 2020

implementation.[261] However, as the literature shows,[262] it might happen that an algorithm cannot be totally purged of all different types of biases. You should, however, try to at least be aware of their existence and the implications that this might bring (see "Lawfulness, fairness and transparency principle" within Part II section "Principles" of these Guidelines).

4.2.3 Exercising data subjects' rights

Sometimes, developers complete the available data through inference. For instance, if you do not have the factual data corresponding to the political opinions of an offender, you might use another algorithm to infer it from the rest of the data, like observed participation in demonstrations. However, this does by no means mean that these data can be considered as pseudonymized or anonymized. Thus, they continue to be personal data. Correspondingly, inferred data must also be regarded as personal data. Therefore, data subjects have some fundamental rights on these data that you must respect.

Indeed, you must respect data subjects' rights during the whole life cycle. In this specific stage, right to access, rectification and erasure are particularly sensitive and include certain characteristics that controllers need to be aware of. However, in the case of research for scientific purposes such as the one you are developing, the GDPR includes some safeguards and derogations relating to processing (Art. 89). You must be aware of the concrete regulation in your Member state. According to the GDPR, Union or Member State law may provide for derogations from the main rights included in articles 15 and ff. in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes.

-Right of access (see "Right to access" within Part II section "Data subject rights" of these Guidelines)

In principle, you shall respond to data subjects' requests to gain access to their personal data, assuming they have taken reasonable measures to verify the identity of the data subject, and no other exceptions apply. However, you do not have to collect or maintain additional personal data to enable the identification of data subjects in training data for the sole purposes of complying with the regulation. If you cannot identify a data subject in the training data and the data subject cannot provide additional information that would enable their identification, they are not obliged to fulfil a request that is not possible to satisfy.

-Right to rectification (see "Right to rectification" within Part II section "Data subject's rights" of these Guidelines)

[261] https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf Accessed 15 May 2020

[262] Chouldechova. Alexandra, Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments, Big Data. Volume: 5 Issue 2: June 1, 2017. 153-163.http://doi.org/10.1089/big.2016.0047

In the case of the right to rectification, you must guarantee the right to rectification of the data, especially those generated by the inferences and profiles drawn up by an AI tool. Even though the purpose of training data is to train models based on general patterns in large datasets, and thus individual inaccuracies are less likely to have any direct effect on a data subject, the right to rectification cannot be limited. As a maximum, you could ask for a longer period (two extra months) to proceed with the rectification if the technical procedure is particularly complex (Art. 11(3)).

-Right to erasure (see "Right to Erasure" within Part II section "Data subject's rights" of these Guidelines)

Data subjects hold a right to delete their personal data. However, this right might be limited if some concrete circumstances apply. According to the British ICO, "organizations may also receive requests for erasure of training data. Organizations must respond to requests for erasure, unless a relevant exemption applies and provided the data subject has appropriate grounds. For example, if the training data is no longer needed because the ML model has already been trained, the organization must fulfil the request. However, in some cases, where the development of the system is ongoing, it may still be necessary to retain training data for the purposes of re-training, refining and evaluating an AI tool. In this case, the organization should take a case-by-case approach to determining whether it can fulfil requests. Complying with a request to delete training data would not entail erasing any ML models based on such data, unless the models themselves contain that data or can be used to infer it."[263]

5 Evaluation (validation)

5.1 Description

"Before proceeding to final deployment of the model built by the data analyst, it is important to more thoroughly evaluate the model and review the model's construction to be certain it properly achieves the business objectives. Here it is critical to determine if some important business issue has not been sufficiently considered. At the end of this phase, the project leader then should decide exactly how to use the data mining results. The key steps here are the evaluation of results, the process review, and the determination of next steps."[264]

[263] ICO, Enabling access, erasure, and rectification rights in AI tools, At: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/ Accessed 15 May 2020

[264] Colin Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, p. 17

This phase involves several tasks that raise important data protection issues. Overall, you must:

- Evaluate the results of your model, for instance, whether it is accurate or not. To this purpose, the AI developer might test it in the real world. This test could often be done in coordination with a project related partner from the domain the system should be rolled-out (e.g. LEA)
- Review the process. You shall review the data processing system to determine if there is any critical factor or task that has somehow been overlooked. This includes quality assurance issues. This actually is the absolute latest phase to involve potential end-users into the development process. However, you should involve and learn about the needs of the end-user in a really early stage of your project (Business understanding). At this stage, stakeholders and end-users can give insights into the system's strengths and weaknesses in real-world use.

5.2 Main actions that need to be addressed

5.2.1 Processes of dynamic validation

The validation of the processing, including an AI component, must be done in conditions that reflect the real environment in which the processing is intended to be deployed. Thus, if you know in advance where the AI tool will be used, you should adapt the validation process to that environment. This is best done by involving respective partners from the domain at stake. If the tool will be deployed in country x, you should validate it with data obtained from the respective population or, if not possible, a similar one. Otherwise, the results might be utterly incorrect. In any case, you should advise about the conditions of the validation to any possible user.

Moreover, the validation process requires periodic review if conditions change or if there is a suspicion that the solution itself may be altered. For instance, if the algorithm is being fed with data from a specific group of people, you should assess whether or not this changes its accuracy in another part of the population. You must make sure that validation reflects the conditions in which the algorithm has been validated accurately.

In order to reach this aim, validation should include all components of an AI tool, including data, pre-trained models, environments and the behavior of the system as a whole. Validation should also be performed as soon as possible. Overall, it must be ensured that the outputs or actions are consistent with the results of the preceding processes, comparing them to the previously defined policies to ensure that they are not violated.[265] Validation sometimes needs gathering new personal data. In other cases, controllers use data for purposes other than the original ones. In all these cases, controllers should ensure compliance with the GDPR (see "Purpose limitation" section

[265] High-Level Expert Group on AI, Ethics guidelines for trustworthy AI, 2019, p. 22. At: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai Accessed 15 May 2020

in “Principles” and “Data protection and scientific research” in “Main Concepts”, Part II of these Guidelines).

5.2.2 Deleting unnecessary dataset

Quite often, the validation and training processes are somehow linked. If the validation recommends improvements in the model, training should be performed again. Once the AI tool has finally been achieved, the training stage of the AI tool is completed. At that moment, you should implement the removal of the dataset used for this purpose, unless there is a legal need to maintain them for the purpose of refining or evaluating the system, or for other purposes compatible with those for which they were collected in accordance with the conditions of Article 6(4) of the GDPR (see “Define data storage adequate policies” section).

In the event that data subjects request its deletion, you shall have to adopt a case-by-case approach taking into account any limitations to this right provided by the Regulation (see Art. 17(3)).[266]

5.2.3 Performing external audit of data processing

Since the risks of the system you are developing are high, **an audit of the system by an independent third party must be involved**. A variety of different audits can be used. These might be internal or external; they might cover the final product only or be performed with less evolved prototypes. They could be considered a form of monitoring and a transparency tool, which is supposed to be a quality feature as well.

In terms of legal accuracy, AI solutions must be audited to see whether they work well with the GDPR considering a wide range of issues. The High-Level Expert Group on AI stated that *“testing processes should be designed and performed by as diverse group of people as possible. Multiple metrics should be developed to cover the categories that are being tested for different perspectives. Adversarial testing by trusted and diverse “red teams” deliberately attempting to “break” the system to find vulnerabilities, and “bug bounties” that incentivize outsiders to detect and responsibly report system errors and weaknesses, can be considered.”*[267] The auditing must also comprise the fulfilment of the principle of explicability. “*The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.”*[268] In view of the very severe consequences for individuals being suspected or convicted of criminal activities, the applied ML technologies must allow for explicability, among further measures required, so that the developed systems respect fundamental rights. The audit should also focus on the measures implemented to avoid bias, obscurity, hidden profiling, etc., and the correct use of tools such as the DPIA, which can be performed multiple times. Implementing

[266] AEPD, Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción, 2020, p.26. At: https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf

[267] High-Level Expert Group on AI, Ethics guidelines for trustworthy AI, 2019, p. 22. At: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai Accessed 15 May 2020

[268] Ibidem, p.15

adequate data protection policies from the first stages of the lifecycle of the tool is the best way to avoid data protection issues.

6 Deployment

6.1 Description

"Deployment is the process of getting an IT system to be operational in its environment, including installation, configuration, running, testing, and making necessary changes. Deployment is usually not done by the developers of a system but by the IT team of the customer. Nevertheless, even if this is the case, developers will have a responsibility to supply the customer with sufficient information for successful deployment of the model. This will normally include a (generic) deployment plan, with necessary steps for successful deployment and how to perform them, and a (generic) monitoring and maintenance plan for maintenance of the system, and for monitoring the deployment and correct usage of data mining results."[269]

6.2 Main actions that need to be addressed

6.2.1 General remarks

Once you have created your algorithm, you face an important issue. It might happen that it incorporates personal data, openly or in a hidden way. You must perform a formal evaluation assessing which personal data from the data subjects could be identifiable. This can be complicated at times. For example, some AI solutions, such as Vector Support Machines (VSM), could contain examples of the training data by design within the logic of the model. In other cases, patterns may be found in the model that identify a unique individual. In all of these cases, unauthorized parties may be able to recover elements of the training data, or infer who was in it, by analyzing the way the model behaves. If you know or suspect that the AI tool contains personal data (see "Purchasing or promoting access to a database" section in "Main Tools and Actions", Part II), you should:

- Delete them or, on the contrary, to justify the impossibility of doing so, completely or partly because of the degradation it would mean for the model (see "Storage limitation" section in "Principles" chapter).
- Determine the legal basis for carrying out the communication of personal data to third parties, especially if special categories of data are involved (see "Lawfulness" subsection in "Lawfulness, fairness and transparency principle" within Part II section "Principles" of these Guidelines).
- Inform the data subjects of the processing above.

[269] SHERPA, Guidelines for the Ethical Development of AI and Big Data Systems: An Ethics by Design approach, 2020, p 13. At: https://www.project-sherpa.eu/wp-content/uploads/2019/12/development-final.pdf Accessed 15 May 2020

- Demonstrate that the data protection by design and by default policies have been implemented (especially data minimization) (see "Data protection by design and by default" within Part II section "Main concepts" of these Guidelines).
- Conduct a Data Protection Impact Assessment (DPIA) (see "DPIA" within Part II section "Main Tools and Actions" of these Guidelines)

Finally, you must take regular action to proactively evaluate the likelihood of the possibility of personal data being inferred from models in light of the state-of-the-art technology, so that the risk of accidental disclosure is minimized. If these actions reveal a substantial possibility of data disclosure, necessary measures to avoid it should be implemented (see "Integrity and confidentiality principle" within Part II section "Principles" of these Guidelines).

6.2.2 Updating information

If the algorithm is implemented by a third party, you must communicate the results of the validation and monitoring system employed at the development stages and offer your collaboration to continue monitoring the validation of the results. It would also be advisable to establish this kind of coordination with third parties from whom you acquire databases or any other relevant component in the life cycle of the system. If this involves data processing by a third party, you must ensure that access is provided on a legal basis.

It is necessary to offer real-time information to the end-user about the values of accuracy and/or quality of the inferred information at each stage (see "Accuracy principle" within Part II section "Principles" of these Guidelines). When the inferred information does not reach minimum quality thresholds, you must highlight that this information has no value. This requirement often implies that you shall provide detailed information about the training and validation stages. Information about the datasets used for those purposes is particularly important. Otherwise, the use of the solution might bring disappointing results to the end-users, who are left speculating on the cause.

You must also ensure that any real-world implementation also complies with the *Data Protection Law Enforcement Directive* (Directive 2016/680)[270] and their specific implementation in individual member states. Please be aware that this usually implies for LEAs less restrictive regulations concerning the use of personal data. In criminal justice, the provision of evidence is often a burdensome activity. It is therefore a natural tendency to collect and process as much data as possible that eventually could prove as useful. This tendency is even reinforced by the increasing technical possibilities to analyze huge amounts of data automatically by AI tools. However, data minimization is necessary and effective countermeasures against extensive data collection and processing therefore must be integrated already into the design of the AI tools.

[270] European Parliament and the Council, 2016, DIRECTIVE (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, Official Journal < http://eur-lex.europa.eu/legal-content/EL/TXT/?uri=OJ:L:2016:119:TOC >.

Compliance with human rights and ethical principles requires the fulfilment of further essential conditions:

"Regarding surveillance technologies, the burden of proof should lie with states and/or companies, who have to demonstrate publicly and transparently, before introducing surveillance options,

- that they are necessary

- that they are effective

- that they respect proportionality (e.g. purpose limitation)

- that there are no better alternatives that could replace these surveillance technologies

These criteria must then also be subjected to post factum assessment, either on the level of normal political analysis, or through Member States policies to do so."[271]

[271] European Group on Ethics in Science and New Technologies. (2014). Opinion No. 28: Ethics of security and surveillance technologies (10.2796/22379). Retrieved from Luxembourg: Brussels: https://publications.europa.eu/en/publication-detail/-/publication/6f1b3ce0-2810-4926-b185-54fc3225c969/language-en/format-PDF/source-77404258