AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

D4.3

Initial analysis of the legal and ethical framework of trusted AI

Project Title	AI4Media - A European Excellence Centre for Media, Society and Democracy
Contract No.	951911
Instrument	Research and Innovation Action
Thematic Priority	H2020-EU.2.1.1. - INDUSTRIAL LEADERSHIP - Leadership in enabling and industrial technologies - Information and Communication Technologies (ICT) / ICT-48-2020 - Towards a vibrant European network of AI excellence centres
Start of Project	1 September 2020
Duration	48 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Deliverable title	Initial analysis of the legal and ethical framework of trusted AI
Deliverable number	D4.3
Deliverable version	1.0
Previous version(s)	-
Contractual date of delivery	28 February 2022
Actual date of delivery	25 February 2022
Deliverable filename	AI4Media_D4.3_final.docx
Nature of deliverable	Report
Dissemination level	Public
Number of pages	77
Work Package	WP4
Task(s)	T4.1
Partner responsible	KUL
Author(s)	Lidia Dutkiewicz (KUL), Noémie Krack (KUL), Emine Ozge Yildirim (KUL), Mara Graziani (HES-SO)
Editor	Aleksandra Kuczerawy (KUL), Lidia Dutkiewicz (KUL), Noémie Krack (KUL), Emine Ozge Yildirim (KUL)
EC Project Officer	Evangelia Markidou

Abstract	The initial version of the analysis of the legal and ethical framework of trusted AI provides an overview on how the GDPR provisions should be interpreted when applied in an AI system context. The analysis interprets the EU data protection framework relevant for the AI systems considering the guidelines, opinions and scholarly literature on the subject. Gaps, unclarities and challenges related to the applicability of the data protection provisions to the AI systems are being identified. Then, the deliverable reflects on the upcoming EU legislation which may have an impact on the data protection when applied to AI. Finally, the deliverable provides initial ways forward to reconcile AI and the GDPR and solve the challenges identified.

Keywords	AI, Ethics, Data Protection, GDPR, Transparency, Accuracy, Data Minimisation, Data subject's rights, Lawful basis, Storage limitation, Purpose limitation, Right of access, Right to rectification, Right to be forgotten, Right to object, Right to be informed, AI Act, Data Governance Act, Recommendations, Standards, Codes of conduct.

Copyright

Contributors

NAME	ORGANISATION
LIDIA DUTKIEWICZ	**KUL**
NOÉMIE KRACK	**KUL**
EMINE OZGE YILDIRIM	**KUL**
ALEKSANDRA KUCZERAWY	**KUL**
MARA GRAZIANI	**HES-SO**
HENNING MÜLLER	**HES-SO**

Peer Reviews

NAME	ORGANISATION
THOMAS KÖLMER	**FhG-IDMT**
REMI MIGNOT	**IRCAM**
FILARETI TSALAKANIDOU	**CERTH**

Revision History

VERSION	DATE	REVIEWER	MODIFICATIONS
0.1	**16/02/2022**	**LIDIA DUTKIEWICZ, NOEMIE KRACK, EMINE OZGE YILDIRIM (KUL)**	**Draft version sent to internal reviewers**
0.2	**25/02/2022**	**THOMAS KÖLMER (FhG-IDMT), REMI MIGNOT (IRCAM), FILARETI TSALAKANIDOU (CERTH)**	**Reviewers' comments implemented**
1.0	**25/02/2022**	**LIDIA DUTKIEWICZ, NOEMIE KRACK (KUL)**	**Final version ready for submission**

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Table of Abbreviations and Acronyms

Abbreviation	Meaning
AI	Artificial Intelligence
AI HLEG	High-level Expert on Artificial Intelligence
API	Application Programming Interface
Art.	Article
ATAP	Algorithmic Transparency and Accountability in Practice
B2B	Business to Business
D.	Deliverable
DGA	Data Governance Act
DPA	Data Protection Authority
DPIA	Data Protection Impact Assessment
EC	European Commission
EDPB	European Data Protection Board
EDPS	European Data Protection Supervisor
EPRS	European Parliamentary Research Service
et al.	And others
EU	European Union
GDPR	General Data Protection Regulation
ICO	Information Commissioner's Office
IFTTT	IF This Then That
ML	Machine Learning
Rec.	Recital
TCF	Transparency and Consent Framework
WP	Working Package
WP29	Article 29 Data Protection Working Party
XAI	Explainable AI

Index of Contents

Index of Tables

Index of Figures

1. Executive Summary

This Deliverable 4.3 "*Initial analysis of the legal and ethical framework for trusted AI*" provides an analysis of whether and when the General Data Protection Regulation (GDPR) provisions apply to AI systems. Then, the analysis addresses the question of how the GDPR provisions should be interpreted when applied in an AI system context, in order to move towards an ethical and legal framework for trusted AI.

Section 2 presents the various objectives of the deliverables and defines the key concept which will be used throughout the document.

Section 3 constitutes the core part of the research conducted for this deliverable. The section shows that despite the GDPR not referring to 'artificial intelligence' many provisions of the legal text prove to be relevant for AI systems. The section first introduces the use of AI systems in the media environment, including recommender and targeted advertising systems. Then **sub-section 3.1** studies the GDPR principles one by one and how the AI systems considerations can be associated with the various principles. This sub-section addresses the principles of lawfulness, fairness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality, and accountability. **Sub-section 3.2** provides an analysis of the different data subjects' rights when applied in an AI context. The following rights are considered: the right to be informed, the right not to be subject to a decision based solely on automated processing, the so-called right to explanations, the right of access, the right to rectification, the right to erasure, the right to restrict processing and the right to object. Then, **sub-section 3.3** addresses the challenges to comply with the data subject's request for rights enforcement in big data sets. The challenges vary from complexities related to the different stages of AI processing to the transparency problems. It also includes a lack of friendly interfaces and technical tools to enforce data subjects' rights. The issues with GDPR enforcement also contribute to trade-offs strategies by companies. This harms considerably the enforcement of data subject's request and the protection of the rights. The absence of explicit reference and further explanations on how the GDPR concepts could be applied to an AI system environment create a need for further guidance on the topic.

Section 4 focuses on upcoming European legislations which appear to be relevant for GDPR provisions and AI systems processing personal data. It considers the AI act proposal (**sub-section 4.1**), the Data Governance Act proposal **(sub-section 4.2)** and the forthcoming Data Act proposal (**sub-section 4.3**).

Section 5 provides initial recommendations for trusted and GDPR-compliant AI. **Sub-section 5.1** offers a conclusion on the gaps and challenges identified throughout the deliverable. **Subsection 5.2** provides ways forward to mitigate and prevent the issues for trusted AI.

2. Introduction

In Section 2.1, we introduce the reader to the purpose of the deliverable and the aim behind the initial analysis of the legal and ethical framework of trusted AI. The notions used through the deliverable will be further conceptualised in Section 2.2. This includes data processing, personal data, data pseudonymisation and anonymisation. In addition, the different roles organisations can play under the GDPR in an AI context are detailed.

2.1 The purpose of this document

The purpose of this deliverable is to:

1) provide a general introduction into the General Data Protection Regulation (GDPR) principles and their applicability to AI (at the collection, training, deployment stage, whenever applicable). This analysis will shed some light on potential gaps, unclarities preventing trusted AI to develop.

2) reflect on the upcoming EU legislation which may have an impact on the (re-)use of data for AI operations;

3) present initial observations about measures to solve the challenges identified and ways forward towards trusted AI.

This document does not aim to provide legal guidance. It also does not aim to replace the existing documents on AI and the GDPR. Among many useful documents from scholars, the following ones are worth mentioning and could be considered as points of reference to any GDPR and AI-related questions: '*Artificial intelligence and data protection: an exploratory guide'* by Knowledge Centre Data & Society, or '*The impact of the General Data Protection Regulation (GDPR) on artificial intelligence'* by the European Parliamentary Research Service (EPRS).

We rather aim to provide explanations of the GDPR principles in an accessible language, and, whenever possible, offer hands-on, practical examples of their implementation.

2.2 The notions used

Throughout this document, we will use language which is being used by the GDPR. It is therefore crucial to understand what these legal notions mean.

➢ What is personal data "processing"?

The term *'processing'* is very broad. It means any operation or set of operations performed on personal data or on sets of personal data, whether or not by automated means. It includes, but is not limited to: collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction. In other words, personal data is processed as soon as something is done with this data or even as soon as the

data passes through an environment controlled by the organisation, even if there is no effective access and the organisation does not do anything else with the personal data.

- What are “personal data”?

It is crucial to know what kind of data is or will be used in a given AI system. **If the system processes personal data, then the GDPR must be complied with.**

Personal data are both data that make it possible to identify a natural person and data that relate to an identified or identifiable person. An individual is *identified* when, within a group of persons, she is distinguished from all other members of the group. Individuals can be identified by e.g., name or address ('direct identification'), but also by their IP address, cookie identifier, location data, or other factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person ('indirect identification'). If a person cannot be immediately identified, it must be verified whether indirect identification is possible or not.

On the other hand, *identifiable* means that, although the person has not been identified yet, it is still possible to do so. To ascertain whether an individual is identifiable, Rec. 26 of the GDPR specifies that ‘*account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly*’. Whether the means are ‘reasonably likely’ must be assessed in light of ‘*objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments*’.

This means that establishing the identifiability of the person, and consequently the applicability of the GDPR, requires a dynamic, context-sensitive analysis of the factual situation. Thus, the exact same dataset might be considered as not containing personal data at the start of the processing and, later on, it might fall under the definition of ‘personal data’ given the tools and data available to the data controller. The same might happen depending on who is actually processing the datasets.

- What are special categories of personal data?

Special categories of personal data (also commonly called sensitive personal data) are:

- personal data revealing racial or ethnic origin;
- personal data revealing political opinions;
- personal data revealing religious or philosophical beliefs;
- personal data revealing trade union membership;
- genetic data;
- biometric data (where used for identification purposes);
- data concerning health;
- data concerning a person’s sex life; and
- data concerning a person’s sexual orientation.

You must always ensure that the data processing is generally lawful, fair and transparent and complies with all the other principles and requirements of the GDPR. To process personal data,

you must always fulfil one of the lawful bases of Article 6 of the GDPR (see Section 3.1.1.1). In addition, you can only process special category data if you can meet one of the specific conditions in Article 9 of the GDPR.

- What is the significance of the distinction between anonymisation and pseudonymisation?

A major distinction has to be drawn between pseudonymized and anonymized data.

Pseudonymization is defined in Art. 4(5) of the GDPR as '*the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person'.*

From this definition, two things come to the fore: first, pseudonymization is a personal data processing activity and is, as such, subject to the GDPR. Second, pseudonymized data resulting from the pseudonymisation activities are still personal data and remain subject to the GDPR.

A further distinction shall be made here, with respect to how the "disguise" of the identify of data subjects is conducted. This can be done in:

- a retraceable way, e.g., using correspondence lists or two-way cryptography algorithms, or
- in a non-retraceable way, e.g., using one-way cryptography algorithms.

In the first case, individuals are still indirectly identifiable since it is possible to backtrack their identity using additional information. Resulting data are "pseudonymized data" and thus still considered personal data.

In the second case, individuals are no longer identifiable since the link between their pseudonym and identity is either inexistent or has been permanently deleted. Such non-retraceable pseudonymization techniques generally create anonymized data that are not subject to data protection rules. The key criterion in distinguishing pseudonymized data from anonymized data is whether individuals are identifiable. This calls for a case-by-case analysis of the factual circumstances surrounding the processing operations.

Anonymous information is defined in Rec. 26 of the GDPR as *"information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable".* In such case, the GDPR is not applicable to the processing of such data.

The creation of a truly anonymized dataset from personal data whilst not depriving the information it carries from its added value is not a trivial task. Depending on the technical possibility and risks of re-identification, data may sometimes still be considered as personal from a legal perspective. The Article 29 Working Party (WP29, currently European Data Protection Board, EDPB) highlights that in determining whether or not the data are still identifiable, focus should be placed on the concrete means that would be necessary to reverse the anonymization technique, particularly the knowledge how to implement those means and the assessment of their likelihood and severity (Article 29 Data Protection Working Party 2014, Opinion 05/2014 on Anonymisation Techniques). For example, encrypted personal data will be anonymous data,

when it would require an excessively high effort or cost or it would cause serious disadvantages to reverse the process and re-identify the individual.

Additionally, one must bear in mind that the means to be assessed are not only those of the data controller, but also the ones that may be used by any other person. **True anonymization is consequently a very onerous standard, and the notion calls for vigilance when used.** In particular, having gone through anonymization process at a certain point in time should not be viewed as a silver bullet for circumventing the application of the GDPR, as identification of natural persons may happen in further processing activities (e.g., when aggregating such data with other data). In addition, critical views were expressed on the efficiency of anonymisation in a big data world, researchers showed how even anonymised datasets can be traced back to individuals using machine learning (Rocher, Hendrickx, and de Montjoye 2019). They demonstrated that allowing data to be used to train AI algorithms would require much more work than simply adding noise, sampling datasets, and other de-identification techniques.

- What about mixed datasets with personal and non-personal data?

According to the EC Guidance (European Commission, 2019), in the case of a mixed dataset where personal and non-personal data are 'inextricably linked', "*the data protection rights and obligations stemming from the GDPR fully apply to the whole mixed dataset, also when personal data represent only a small part of the dataset".* While it is not entirely clear what '*inextricably linked'* means, it is safe to assume that the GDPR is fully applicable to such datasets.

- What are the different roles an organisation can play under the GDPR in an AI context?

One of the most important aspects under the GDPR is defining the different roles and responsibilities with regard to the processing of personal data. Throughout this deliverable we will be referring to **data controllers** and **data processors**. The distinction between (**data**) **controller** and (**data**) **processor** is important, because they each have different obligations under the GDPR.

At the different stages of the life cycle of an AI system, the **controller** is the natural or legal person, public authority or other organisation that **decides on the purposes and means** of processing personal data.

Processor, on the other hand, means a natural or legal person, public authority, agency or other body which processes personal data **on behalf of** the controller.

Moreover, if two organisations jointly determine the purposes and means of the processing through an AI system, they may be considered as **joint controllers**. This may be the case, for example, where an organisation cooperates with another organisation in developing a product or service for which both parties provide personal data for the training and/or validation of the tool, and where they jointly determine the purpose of such processing and combine their technical resources, without one party processing personal data solely on the instructions of the other. The Table 1 below provides a schematic overview of different roles a data controller and data processor can have in an AI context.

Table 1: Different roles an organisation can play under the GDPR in an AI context

PHASE/ACTIVITY	WHO IS THE CONTROLLER?	WHO IS THE PROCESSOR?
DEVELOPMENT/ TRAINING/ VALIDATION	The person or organisation that (further) develops, trains or validates the AI system and decides what personal data will be used to train the system (and therefore determines the purpose and means). If this organisation obtains a set of personal data from a third party, it will also have the status of controller when processing such data. If the development, training, validation or (further) development is outsourced to a third party organisation and this third party organisation decides which type of personal data is used in this regard, it becomes a controller.	The organisation to which the development, training, validation or (further) development is outsourced, provided that the client to whom such services are provided: (i) identifies the purpose of the processing activity and; (ii) determines the significant characteristics of the personal data to be processed. This is regardless of whether this client/controller transfers the personal data to the processor or the processor obtains it through its own channels and; (iii) the processor processes such data only for the purposes specified by the controller.
LAUNCH/ RELEASE/ COMMISSIONING	Any organisation that integrates an AI system into its product or service and thereby processes personal data for its own purposes. If the AI system (whether or not part of a wider product or service) is sold or licensed and already contains personal data, both organisations exchange personal data and are both controllers. Even if, for instance, a licensor makes a system available to a licensee and only the licensee is the controller (see on the right), the licensor still also becomes a controller when it processes personal data obtained from the licensee for its own purposes (e.g., to	Any organisation that makes an AI system available to a controller whereby the AI system is integrated into the latter's product or service, or any organisation that does so because it is necessary for the proper performance of its service, but that does not itself process personal data obtained from the controller for its own purposes. An organisation (service provider) that makes an AI system available to another organisation (user) is neither a processor nor a controller if: (i) this system is installed locally and stand-alone at the user's premises;

	measure the efficiency of the AI system).	(ii) the service provider does not have access to the local installation, e.g., for maintenance.

3. AI and the GDPR

The GDPR does not contain the term 'artificial intelligence', nor any terms expressing related concepts, such as intelligent systems, autonomous systems, automated reasoning and inference, machine learning or even big data. This does not, however, mean that the GDPR does not apply to training, testing, validation or deploying the AI systems. To the contrary, as presented in Figure 1, many provisions in the GDPR are very relevant to AI.

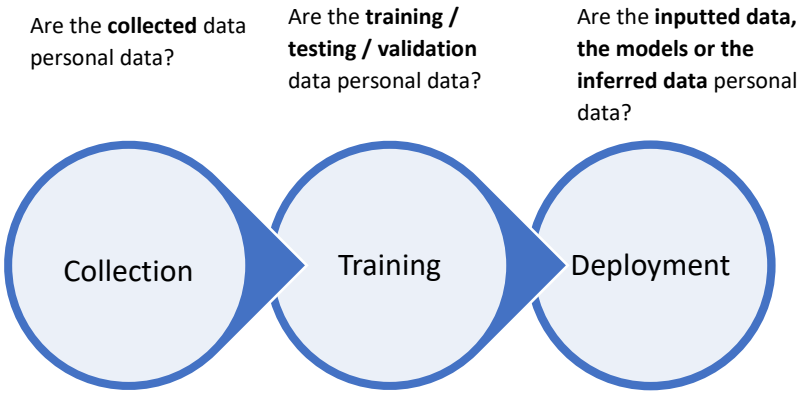

Figure 1: When does the GDPR apply to AI operations?

Generally speaking, many AI applications process personal data. On the one hand, personal data may contribute to the data sets used to train ML systems, namely, to build their algorithmic models. On the other hand, such models can be applied to personal data, to make inferences concerning particular individuals. Next, thanks to AI processing, personal data can be used to analyse, forecast and influence human behaviour. In the context of media, personal data are often collected, processed and used for many purposes, among which automated personalisation of (recommendations for) content (e.g., news) and advertising (e.g., targeted advertisements).

Example 1: Automated data capture and processing

Many digital media, platforms and websites use a variety of AI technologies to capture and process data for reasons of personalisation, profiling, inferential predictive analytics, targeted advertising, etc. Given that the advertising-driven business model is a significant part of digital media, special attention is needed on how data is captured and processed with regard to advertisement technology and marketing automation. This is particularly relevant for online behavioural advertising, where internet users' behavioural data (website visits, clicks, mouse movements, etc.) and metadata (browser type, location, IP address, etc.) are collected and processed to create profiles used to personalise ads (Pierson et al 2021). The profiles contain categories of users' past behaviour, but also inferred preferences and affinities, being often sensitive categories of data protected by the GDPR. For example, Google and several data brokers have been accused of violating the GDPR rules by harvesting and processing people's personal data to build detailed online profiles, including information on sexual orientation, health status and religious beliefs (Scott and Manancourt 2021).

In that regard, in a recent decision of 2 February 2022, in Case No. DOS-2019-01377, 28 EU data protection authorities, led by the Belgian Data Protection Authority (DPA), found that the online advertising industry's trade body "IAB Europe" commits multiple violations of the GDPR in its processing of personal data in the context of its consent popup system (Transparency and Consent Framework, TCF) and the Real-Time Bidding system. In what can be considered a landmark decision with major impact in Europe, the Belgian DPA found that the TCF deprived hundreds of millions of Europeans of their fundamental rights. The TCF consent system was found to infringe the GDPR in the following ways:

- TCF fails to ensure personal data are kept secure and confidential (Article 5(1)f, and 32 of the GDPR)
- TCF fails to properly request consent, and relies on a lawful basis (legitimate interest) that is not permissible because of the severe risk posed by online tracking-based "Real-Time Bidding" advertising (Article 5(1)a, and Article 6 of the GDPR)
- TCF fails to provide transparency about what will happen to people's data (Article 12, 13, and 14 of the GDPR)
- TCF fails to implement measures to ensure that data processing is performed in accordance with the GDPR (Article 24 of the GDPR)
- TCF fails to respect the requirement for data protection by design (Article 25 of the GDPR)

All data collected through the TCF must now be deleted by the more than 1,000 companies that pay IAB Europe to use the TCF. This includes Google's, Amazon's and Microsoft's online advertising businesses.

Example 2: Automated content mediation
Automated content mediation involves automated filtering systems in the distribution and moderation of online content ('algorithmic content moderation') and advertising. AI technologies in content distribution occur in the form of recommender systems, online news aggregators, and programmatic advertising, which provide user-specific content. To monitor and moderate online content, a range of personal and non-personal data must be stored by the company, such as the username of the individual, the name of the complainant, the justification for the removal of the content, dates and times of uploads and removals and so on.

Example 3: Automated communication
Automated communication, includes AI-enabled communication infrastructure such as chatbots, smart speakers, virtual voice assistants and so on. All these tools work with user personal data as they 'learn' from this data. This includes primary data (e.g. account data, voice recordings, requests history), observed data (e.g. device data that relates to a data subject, activity logs, online activities), as well as inferred or derived data (e.g. user profiling). As noted by the EDPB, the personal data processed by virtual voice assistants may be highly sensitive in nature. It may carry personal data both in its content (meaning of the spoken text) and its meta-information (sex or age of the speaker etc.) (EDPB, 2021).

3.1 The GDPR principles

All of the data protection principles apply to personal data processing, but perhaps most significant are the requirements of the first principle: lawfulness, fairness and transparency. In what follows, we will provide the overview of each element of this principle and provide guidance on the practical application of this principle to AI and machine learning. For complementarity, we then describe other GDPR principles.

3.1.1 Lawfulness, fairness and transparency

Article 5(1)(a) of the GDPR requires that personal data should be processed **'lawfully, fairly and in a transparent manner** in relation to the data subject.

3.1.1.1 Lawfulness

According to Article 8 of the Charter of Fundamental Rights of the European Union, personal data must be processed fairly for specified purposes and on the basis of a legitimate basis laid down by law. In this regard, Article 6(1) of the GDPR specifies that data processing is lawful only if it is based on one of six specified conditions set out in Article 6(1)(a) to (f) (see Figure 2 below).

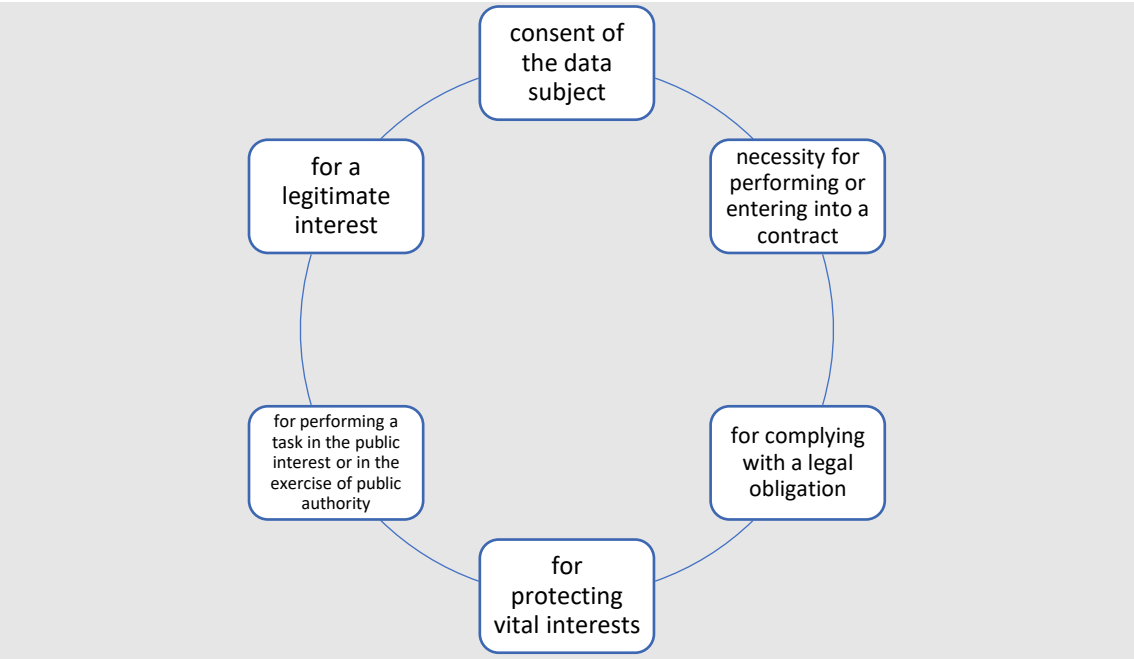

Figure 2: Lawful basis to process personal data

Identifying the appropriate lawful basis is of essential importance. Whenever you are processing personal data – whether to train a new AI system or make predictions using an existing one – you must have an appropriate lawful basis to do so.

- How to identify lawful basis when using AI?

Different lawful bases may apply depending on particular circumstances. However, some lawful bases may be more likely to apply for the training and/or deployment of AI than others. In some cases, more than one lawful basis may be appropriate. Nevertheless, as provided by the ICO's guidance on AI and data protection, the following must be remembered when deciding about the lawful basis for processing personal data:

- it is your responsibility to decide which lawful basis applies to your processing;
- you must always choose the lawful basis that most closely reflects the true nature of your relationship with the individual and the purpose of the processing;
- you should make this determination before you start your processing;
- you should document your decision;
- you cannot swap lawful bases at a later date without good reason;
- you must include your lawful basis in your privacy notice (along with the purposes); and
- if you are processing special categories of data, you need both a lawful basis and an additional condition for processing (ICO 2021).

➢ How to distinguish lawful basis between AI development and deployment?

Development (including conceptualisation, design, training and model selection) and deployment are two different stages of the AI lifecycle. It is not surprising that they may a require different lawful basis. It would be the case, where, for example:

- the AI system was developed for a general-purpose task, and its subsequently deployed in different context for different purpose;
- when implementing an AI system from a third party: the purpose to develop the system is different from what you intend to use the system for, and therefore requires a different lawful basis; and
- processing of personal data for the purposes of training a model may not directly affect the individuals, but once the model is deployed, it may make automated decisions, which have legal or significant effects. This means the provisions on automated decision-making apply (see Section 3.1.1.3 below).

➢ What constitutes a lawful basis?

a) Article 6(1)(a) of the GDPR: Consent

Consent may be an appropriate lawful basis in cases where you have a **direct relationship** with the data subject whose data you want to process.

Article 4(11) of the GDPR defines consent as:

> Consent of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.

You may rely on a data subject's consent to process personal data, to include such data in a training set, to provide them to an AI model or during deployment of an AI system (e.g., for purposes such as personalizing the service or making a prediction or recommendation). However, to rely on consent data subjects must have a genuine choice about whether their data will be used and for what purpose. Their consent must be genuinely specific and informed. In practice, it is not easy to satisfy all the above-mentioned conditions.

Moreover, for consent to be valid, individuals must also be able to easily withdraw consent at any time. In such a case, the data controller is required to return the personal data to the data subject or/and delete the data and terminate data processing activities. Note that withdrawing consent may have serious practical implications: the question rises if a data subject consents to have their data used to train a particular model, and then later withdraws that consent, would the model have to be retrained on new data? As the WP29 has specified, even after consent is withdrawn, all processing that occurred before the withdrawal remains legal (WP29 2018). In practice, once a model is created with a set of training data, that training data can be deleted or modified without affecting the model. Technically, however, some research suggests that

models may retain information about the training data in ways that could allow the discovery of the original data even after training data has been deleted (Papernot et al. 2017).

For these reasons, consent may not be the most applicable legal basis to rely on. Data processing for AI purposes usually needs to rely alternatively or additionally on other legal bases.

b) Article 6(1)(b-e) of the GDPR: Necessity

The legal bases from (b) to (e) all involve establishing the necessity of the processing for a certain aim: (b) performing or entering into a contract, (c) for complying with a legal obligation, (d) protecting vital interests, (e) performing a task in the public interest or in the exercise of public authority. There are some limited cases in which the use of an AI system to process personal data may be a legal obligation (e.g., to audit an AI system to ensure they are compliant with legislation). Similarly, if you use AI as part of the exercise of your official authority, or to perform a task in the public interest set out by law, the necessary processing of personal data may be based on those grounds. It is however very unlikely that these grounds could provide a basis for developing an AI system. WP29 concludes that such legal bases do not, in general, apply to AI-based processing (WP29 2018).

c) Article 6(1)(f) of the GDPR: Legitimate interest

Article 6(1)(f) legal basis is the necessity of the data processing for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject.

The WP29 Opinion 06/2014 on the notion of legitimate interests of the data controller developed in depth guidance on how should factors legitimating the interest of the data controller to process personal data be assessed and balanced with the also legitimate rights and interests of the data subjects.

In short, there are three elements to the legitimate interest lawful basis which a data controller has to comply with (Figure 3).

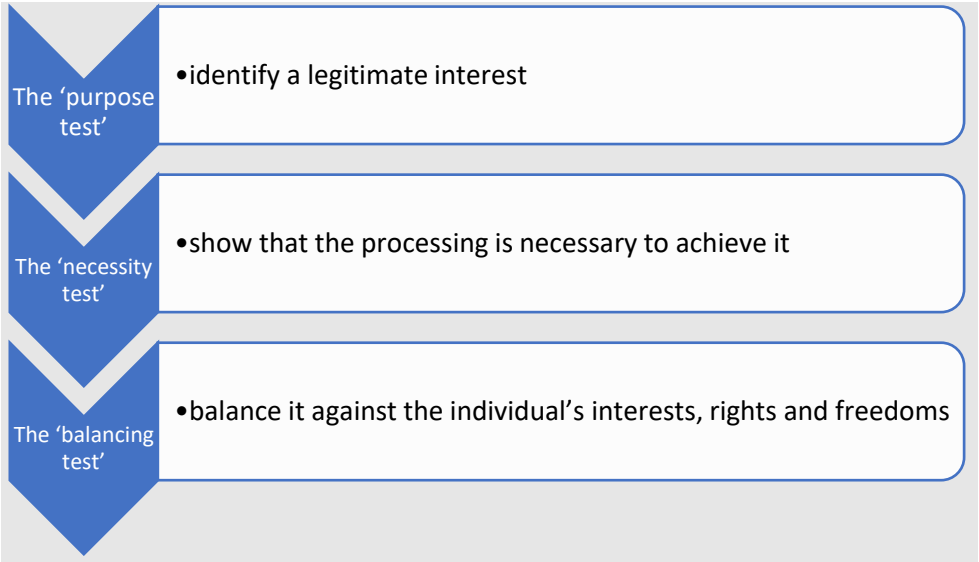

Figure 3: Legitimate interest test

The legitimate interest test requires the data controller to assess the impact of processing on individuals and be able to demonstrate that there is a compelling benefit to the processing.

The ICO's Guidance on AI and Data Protection (ICO 2021) provides the following example:

Example

An organisation seeks to rely on legitimate interests for processing personal data for the purposes of training a machine learning model. Legitimate interests may allow the organisation the most room to experiment with different variables for its model. However, as part of its legitimate interests' assessment, the organisation has to demonstrate that the range of variables and models it intends to use is a reasonable approach to achieving its outcome.

It can best achieve this by properly defining all of its purposes and justifying the use of each type of data collected – this will allow the organisation to work through the necessity and balancing aspects of its Legitimate Impact Assessment (LIA). Over time, as purposes are refined, the LIA is revisited.

For example, the mere possibility that some data might be useful for a prediction is not by itself sufficient for the organisation to demonstrate that processing this data is necessary for building the model.

➢ Article 9 of the GDPR

Article 9 of the GDPR addresses the so-called sensitive or special categories of personal data, such as:

> personal 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

In general, processing of such data is prohibited unless there are exceptional grounds for doing so. The first exception is the explicit consent of individuals. Note that an 'explicit consent' is not defined in the GDPR. An 'explicit' consent is not the same as regular consent (for more information see, for example, Article 29 Working Party Guidelines on consent under Regulation 2016/679).

The second exception is sensitive data which have been 'manifestly made public by the data subject'. 'Manifestly' means that there must be clear evidence of a deliberate, affirmative act by the data subject themselves to make their data available.

Another exception is data processing 'necessary for reasons of substantial public interest, on the basis of Union or Member State law'. A 'substantial' public interest is not the same as 'public interest'. National law must clearly indicate what they mean by such substantial public interest. As provided by the EDPS, in lack of such law *"it is (…) difficult at present, if not impossible, to view a 'substantial public interest' as a basis for processing sensitive data for scientific research purposes"* (EDPS 2020).

Finally, the research exception described in Article 9(2)(j) of the GDPR allows for the processing of sensitive data for research purposes under two conditions. First, there must be EU or national legislation which allows it. Second, if sensitive data (e.g., facial images) are processed for research purposes, the safeguards for the rights and freedoms of the data subject defined in Article 89(1) of the GDPR must be implemented. Those include technical and organizational measures such as anonymization and pseudonymization of personal data.

Importantly, Article 9(4) of the GDPR provides that Member States can adopt stricter requirements for the processing of specific types of sensitive personal data. Those include genetic data, biometric data (e.g., biometric facial images, EEG data), and health-related data. It is not excluded that some Member States require individuals' explicit consent to process for example their biometric data.

➢ Challenge to comply with legal basis: training datasets

AI researchers typically rely on existing datasets as training data, such as CommonCrawl to train large language models, ImageNet for object recognition, or MS COCO for computer vision tasks and object recognition. While not all the data used is personal data (and thus not covered by the GDPR), often these datasets may contain personal data, and at times even special categories of

data. As already said, one must have a lawful basis to process the personal data therein (be that to set up a dataset, re-use it, re-use parts of it or for any other data processing activity).

In 2019, the New York Times revealed that one of these public datasets, MegaFace, used the facial images from an existing photo database set up by Yahoo: the Yahoo Flickr Creative Commons 100 Million (the YFCC100M) Dataset ('How Photos of Your Kids Are Powering Surveillance Technology - The New York Times' n.d.). To assemble MegaFace, researchers have extracted the photos from the Yahoo! Database and matched them, as much as possible, with Flickr accounts. Raji et al. note that *"while these images are open for public internet use, the Flickr users who uploaded the photos, and the individuals in the photos did not consent to being included in a facial recognition dataset."* (Raji et al. 2020). Similarly, after the investigation by the Financial Times, Microsoft's MS Celeb database published in 2016 and containing images of nearly 100,000 individuals was also terminated (despite the fact that "the site was intended for academic purposes")('Microsoft Quietly Deletes Largest Public Face Recognition Data Set | Financial Times' n.d.). The people whose photos were used were not asked for their consent, their images were scraped off the web from search engines. Despite the termination of the MS Celeb and Megaface websites, the datasets still exist in several repositories on GitHub, the hard drives of researchers, cloud services etc. Earlier in 2019, images from the MS Celeb were also repackaged into another facial dataset called Racial Faces in the Wild (RFW).

These revelations sparked an international discussion about the dataset origin, which raises privacy, data protection and liability concerns. Specifically, many ask how ethical it is to use individual's faces without their consent. Besides the ethical concern, there is, however, a pressing legal problem: **the initial collection of personal data** to create the database, which are later re-used for scientific research or other purposes, requires a lawful basis under the GDPR.

When personal data initially collected to set up a database, are further processed for **compatible purposes**, such as scientific research, 'no legal basis *separate from* that which allowed the collection of the personal data is required' (Recital 50 of the GDPR). The question arises as to what extent is it the responsibility of the re-user (e.g., the AI researcher) to investigate the legal basis for the initial data processing (data collection). When it comes to publicly available datasets, it might prove difficult if not impossible to find out whether the initial collection of data was based on a valid GDPR legal basis (e.g., individuals' consent or GDPR-complaint legitimate interest). Should that be the case, the researcher would nonetheless have to find and rely on their own lawful basis under Article 6 of the GDPR to process data. Additionally, when a dataset contains the special categories of data, one of the specific exceptions contained in Article 9(2) of the GDPR must be found. What are the most likely lawful bases?

In the context of **scientific research**, the following legal bases are most likely to apply: the consent of the data subject (Article 6(1)(a) GDPR), the performance of a task carried out in the public interest (Article 6(1)(e) GDPR) or a legitimate interest of the controller or a third party (Article 6(1)(f) GDPR). Yet, there are some issues with those legal grounds. We will briefly run through some of them.

Consent, despite of being ethically desirable, is not likely to provide a sufficient legal basis for various reasons. In case of large-scale datasets with millions of personal data points, some of which may be sensitive data, consent is almost impossible to realise. It is impossible to ask each and every individual for her consent. It is also problematic to contact data subjects to ask for consent in the first place, as one must have a legal basis to do so. Relying on a derived consent from the consent that data subjects gave, for example, in the terms and conditions of a social media platform on which they initially uploaded their personal data, would hardly meet a threshold of 'real' genuine and informed consent as explained in Section a) Article 6 (1) (a) of the GDPR: consent.

When it comes to sensitive data, a way forward would be to rely on Article 9(2)(a) of the GDPR, which allows the processing of sensitive personal data which are 'manifestly made public by the data subject'. It is, however, unlikely that all the personal data that is publicly available (e.g., posted on social media) have been released by the data subject himself or herself (Jasserand 2018). As provided in the EDPS (2020), 'this provision has to be interpreted to imply that the data subject was aware that the respective data will be publicly available which means to everyone' including, in this case, researchers. In case of doubt, 'a narrow interpretation should be applied, as the assumption is that the data subject has voluntarily given up the special protection for sensitive data by making them available to the public including authorities'. The Court of Justice of the European Union might soon clarify this question. The request for a preliminary ruling by the Higher Regional Court of Düsseldorf (C-252/21 - Facebook and Others), concerns, among others, whether visiting social media websites or apps and/or entering information and/or clicking or tapping on the buttons integrated into them by a provider such as Facebook ('Like', 'Share') constitute manifestly making the data public within the meaning of Article 9(2)(e) of the GDPR.

The second legal basis could be the performance of a task carried out in the public interest (Article 6(1)(e) of the GDPR). There is however lack of clear guidance how to interpret the 'public interest', especially in a research context. What constitutes a 'public interest' is a matter of national legislation. In lack of any uniformity in this matter, it is a task of each individual to check whether or how the national legislation defines the 'public interest'. In the case of special categories of data, data processing must moreover be 'necessary for reasons of *substantial* public interest' (Article 9(2)(j) of the GDPR). Again, it is a matter for Member State law to substantiate this provision.

Finally, it can be argued that **the key legal basis for training AI models with personal data will be legitimate interest** under Article 6(1)(f) of the GDPR (Hacker 2021). And although it may hold true for *the training operation* itself, especially when there is a high degree of pseudonymisation, the legitimate interest must be assessed also at the very first stage of data processing activity, namely the collection of personal data to be fed into the ML pipeline.

As provided by the WP29 in its Opinion 6/2014 (2014) the *'interest of carrying out scientific research'*, subject to appropriate safeguards, is one of the '*interests [that] may be compelling and beneficial to society at large'.* The legitimate interest basis of course requires a case-by-case assessment and a balancing test between the interests of the researchers (and society as a

whole) and data subjects' rights and freedoms. Recital 113 of the GDPR seems to weigh decisively in favour of third parties' interest when it states that *'for scientific or historical research purposes or statistical purposes, the legitimate expectations of society for an increase of knowledge should be taken into consideration'.* However, an important hurdle concerns the fact that the basis of legitimate interests is only available to private entities, not public authorities (e.g., public universities) *'in the performance of their tasks'.* This is not to say that public authorities, or those having a hybrid status, depending on the task they perform are *a prori* excluded from the scope of Article 6(1)(e). What matters is how national laws and each intuition status defines what constitutes *'the performance of [their] tasks'.*

3.1.1.2 Fairness

According to the EPRS Study conducted by Sartor (Sartor et al. 2020), two different concepts of fairness can be distinguished in the GDPR. The first, which we may call **'information fairness'** is strictly connected to the idea of transparency. It requires that data subjects are not deceived or misled concerning the processing of their data, as is explicated in Recital 60:

> The principles of fair and transparent processing require that the data subject be informed of the existence of the processing operation and its purposes. The controller should provide the data subject with any further information necessary to ensure fair and transparent processing taking into account the specific circumstances and context in which the personal data are processed. (…) Furthermore, the data subject should be informed of the existence of profiling and the consequences of such profiling.

Informational fairness is also linked to accountability, since it presumes that the information to be provided makes it possible to check for compliance. Informational fairness raises specific issues in connection with AI and big data, because of the complexity of the processing involved in AI applications, the uncertainty of its outcome, and the multiplicity of its purposes.

Recital 71 points to a different dimension of fairness, i.e. what we may call '**substantive fairness'**, which concerns the fairness of the content of an automated inference or decision, under a combination of criteria:

> In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, 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.

It follows that 'fairness' cannot be reduced to a synonym of transparency or lawfulness, but has an independent meaning. Malgieri points out that the idea of fairness can have many possible nuances: non-discrimination, fair balancing, procedural fairness, etc. (Malgieri 2020).

Analyzing the GDPR provisions, Clifford and Ausloos notice that the notion of fairness can have two main meanings: fair balancing and procedural fairness (Clifford and Ausloos 2018). Fair balancing is based on proportionality between data subjects' interests (e.g., the right to privacy, right to data protection) and necessity of purposes of the data controller. Procedural fairness refers to practical implementation of 'fairness' through specific procedures that can improve the level of transparency and lawfulness of a certain data processing in a specific context (Clifford and Ausloos 2018).

Actually, the GDPR does not always describe in details such fair procedures: the data controller is asked to choose and adopt her own procedures in order to make a data processing "fairly transparent" and "fairly lawful", in particular looking at the "specific circumstances and context in which the personal data are processed" (recital 60 and 71) or the "specific processing situations" (Article 6(2) and (3)) (Malgieri 2020).

- Fairness as non-discrimination

The notion of fairness is also often interpreted as non-discrimination. Recital 71 affirms that one should prevent "potential risks" for the interests and rights of the data subject, such as "*discriminatory* effects on natural persons". Similarly, WP29, in its Opinion on Automated Decision-Making, also associates unfairness to discrimination: "profiling may be unfair and create *discrimination*, for example by denying people access to employment opportunities, credit or insurance, or targeting them with excessively risky or costly financial products" (WP29 2018).

- The Myth of Complete AI-Fairness

In 'The Myth of Complete AI-Fairness', Dignum rightly observes that "nothing is ever 100% fair in 100% of the situations, and due to complex networked connection, to ensure fairness for one (group) may lead to unfairness for others" (Dignum 2021). She points out that it is not possible to satisfy some of the expected properties of fairness simultaneously: calibration between groups, balance for false negatives, and balance for false positives. Data calibration means higher levels of false positives and false negatives for some groups. This brings her to conclude that achieving perfectly fair, data-driven, algorithms is an impossible task. Importantly, she adds, the debate should rather focus on the societal and individual impact of false positives and false negatives and on what should be the threshold for acceptation of algorithmic decisions.

- Why fairness cannot be automated

In an influential paper titled "Why fairness cannot be automated", Wachter et al. argue that "automating fairness or non-discrimination in Europe may be impossible because the law, by design, does not provide a static or homogenous framework suited to testing for discrimination in AI systems" (Wachter, Mittelstadt, and Russell 2020).

First, the paper argues that fairness is contextual and cannot (and arguably should not) be automated. This clearly follows from the case law of the Court of Justice of the European Union.

Defining a disadvantaged group(s), legitimate comparator group(s), and evidence of a "particular disadvantage" requires the judiciary to make case-specific choices that reflect local, political, social, and legal context of the case as well as arguments made by both parties. There are very few clear-cut examples of static rules, requirements, or thresholds for defining what constitutes a 'discrimination'. Non-discrimination law is also based on the idea of comparison, another deeply contextual concept. Authors argue that this *'contextual normative flexibility, or 'contextual equality'*, must be respected and facilitated in automated systems (Wachter, Mittelstadt, and Russell 2020).

Second, the authors point out that AI system developers and controllers have very little consistent guidance to draw on in designing considerations of fairness, bias, and non-discrimination into AI and automated systems. This makes the judiciary's approach to 'contextual' "difficult, if not impossible, to replicate in automated systems at scale."

Third, they argue that although the technical community has a vital role to play in providing statistical evidence and in developing tools for detection of bias and measuring fairness, the concept of "contextual equality" needs to be exercised by the judiciary, legislators and regulators. They warn against the situation in which "system developers and controllers alone set normative thresholds for discrimination locally and subjectively without external regulatory or judicial input". Such a situation would undermine Europe's non-discrimination law (Wachter, Mittelstadt, and Russell 2020).

Finally, to reconcile this tension, fairness should not be seen as a problem to be solved through automation or technical fixes alone, but rather requires a collaboration between technical and legal communities.

3.1.1.3 Transparency

- The interdisciplinary perspective on transparency and explainability

The last decade saw a sharp increase in research papers concerning interpretability for AI also referred to as eXplainable AI (XAI). In 2020, the number of papers containing 'interpretable AI', 'explainable AI', 'XAI', 'explainability', or 'interpretability' has increased to more than three times that of 2010. The different perspectives about the technical terminology are discussed in several papers within the specific context of explainable AI and ML design, finding difficult integration within the other domains that are driving and shaping AI development.

Discordance can be noticed on the meaning assigned to the terms by the papers coming from multiple disciplinary domains. Major dividing points emerge on the words:

(i) interpretable and explainable;
(ii) transparency and decomposability;
(iii) intelligible and interpretable.

Diverging definitions are used, in particular, between the technical and the social sciences. Such divergences are illustrated in Figure 4.

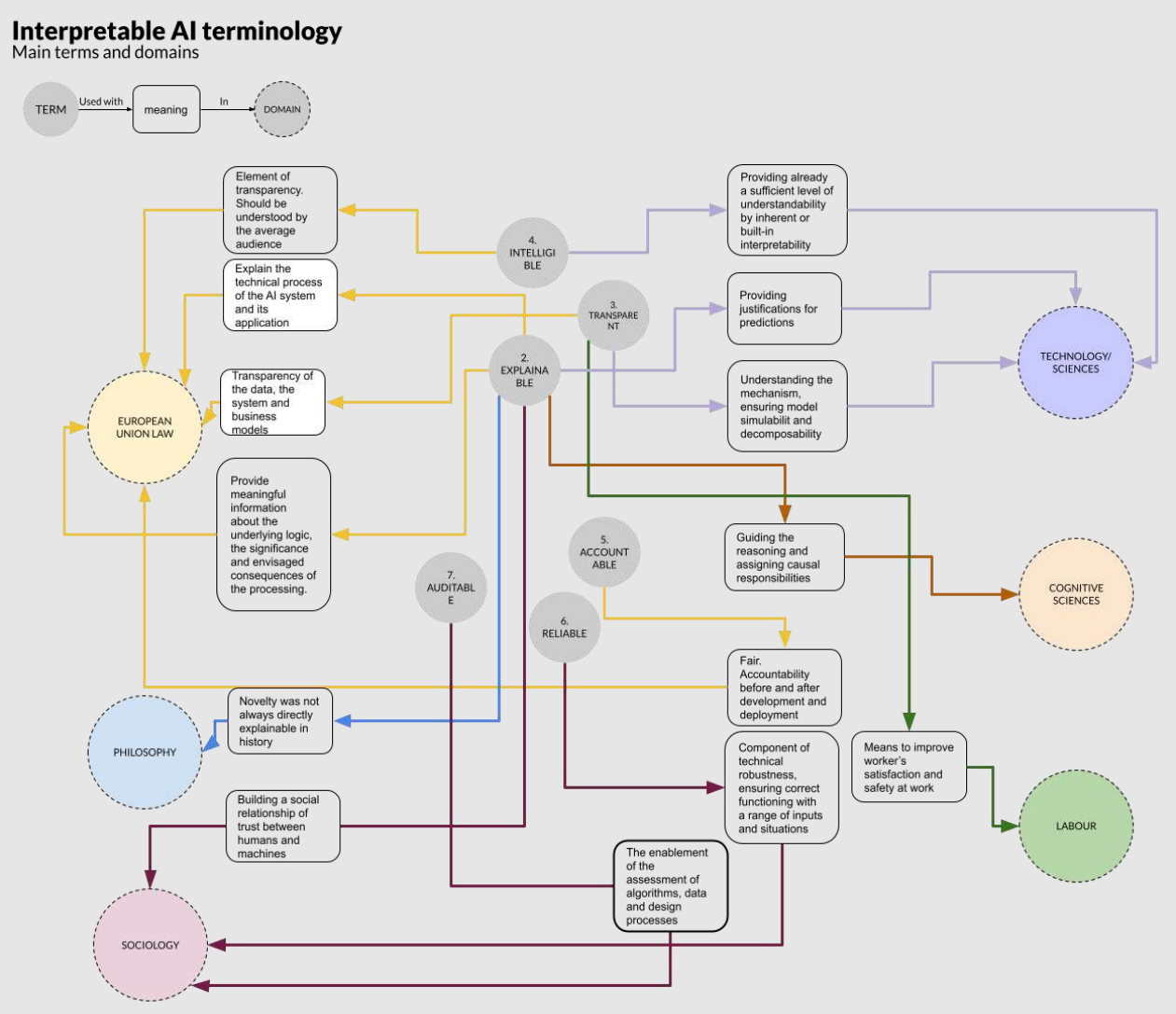

Figure 4: Divergences between the definitions used in social sciences and technical sciences

This analysis is based on the cross-disciplinary knowledge of the people participating in the workshop organised in AI4Media project on 29 April 2019. A round table public meeting held online on "A Global Taxonomy for Interpretable AI" was organized to bring together researchers from multidisciplinary backgrounds to collaborate on a global definition of interpretability that may be used with high versatility in the documentation of social, cognitive, philosophical, ethical and legal concerns about AI. A total of 18 experts were invited to participate in the event. The selection of the experts was tailored to obtain the most representative consortium of the fields dealing with Interpretable AI at the moment. The workshop gave gives insights into how each domain envisions concepts such as transparency, explainability, interpretability etc. Some conflicts in the definitions are shown as the words are used in one or another discipline. The

attention towards one or more concepts is mostly heterogeneous, with some disciplines focusing more on one aspect than others.

While heterogeneity in the attention to the words is legitimate and given by the intrinsic nature of each discipline, the strong changes in the meaning assigned to the same word by different disciplines may inhibit understanding and collaboration among different fields.

As we will see below, the word *transparent* has been interpreted as *"providing meaningful information about the underlying logic"* in the EU legislation, whereas by technical developers this is often understood as a certain degree of understanding of the system mechanics, decomposability and simulability. In other words, if technicians and legislators were to think of the degrees of transparency of a vehicle, they would see to different aspects. The former would think of pistons, fusible and the combination of these elements to the final engine. The latter would think of the degree of information available to the user about the working principles of the vehicle: starting the engine, stopping it from running, changing the direction and so on.

We will now take a closer look on the legal requirements regarding 'transparency' imposed by the GDPR.

➢ General transparency obligations under the GDPR

Articles 12, 13 and 14 of the GDPR contain the main general transparency obligations that controllers must comply with. We briefly discuss them in turn. Other, complementary transparency provisions such as 'data protection by design and by default' (Article 25), records of processing activities (Article 30) and data protection impact assessment (DPIA, Article 30) are not discussed.

a) Article 12 GDPR

Article 12 GDPR lays down in a general manner that the data subject must receive the information in a concise, transparent, intelligible and easily accessible form, using clear and plain language (and where appropriate with visualization; see Recital 39 and 58 of the GDPR). In general, that means that individuals must be informed about the ongoing data processing before such processing takes place.

b) Article 13 and 14 GDPR

Articles 13 and 14 specify in more detail when that information needs to be communicated to a data subject.

In the event of **direct collection** of data from the data subject (Art. 13 of the GDPR), the controller must, at the time when personal data are obtained, provide the data subject with, inter alia, the following information:

- the identity and the contact details of the controller and, where applicable, of the controller's representative;

- the contact details of the data protection officer, where applicable;
- the purposes of the processing for which the personal data are intended as well as the legal basis for the processing;
- the recipients or categories of recipients of the personal data, if any;
- the period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period;
- which rights the data subject has, including the rights a data subject has in the event of automated decision-making (i.e. right to human intervention, right to express an opinion and right to challenge the decision) and how these can be exercised;
- the right to lodge a complaint with a supervisory authority;
- the existence of automated decision-making, including profiling and meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

In case of indirect data collection (Art. 14 of the GDPR), e.g., through a third party, the same information must be communicated, along with:

- the categories of personal data concerned.
- from which source the personal data originate, and if applicable, whether it came from publicly accessible sources.

There are some exceptions to this information obligation in the case of indirect data collection as provided for in Article 14 of the GDPR. For example, this information does not have to be provided in cases where personal data are processed for scientific research purposes and (i) the provision of such information proves impossible or would involve a disproportionate effort, or (ii) the provision of such information is likely to render impossible or seriously impair the achievement of the objectives of that processing. This will be further explained in Section 3.3.

➢ Information requirements specific to AI systems

In Articles 12 to 14 and to a lesser extent in Articles 15 and 22, the GDPR imposes various transparency or information obligations that are crucial for AI applications (Table 2).

Table 2 Information requirements specific to AI systems

GDPR PROVISION	OBLIGATION
Art. 13(2)(f), Art. 14(2)(g), Art. 15(1)(h)[1]	Obligation to inform about **the existence and use** of automated (individual) decision-making and profiling
	Obligation to provide 'meaningful information on the **logic** involved'
	Obligation to inform about the '**significance** and the envisaged **consequences**' of this processing for the data subject

[1] Article 15(1)h is identical to Articles 13(2)f and 14(2)h: data subjects have a right to be informed about the existence of automated decision-making and to obtain meaningful information about the significance, envisaged consequences, and logic involved. However, there is a difference between the information

Art. 22 **Recital 71**	Obligation to provide **explanation** of the individual automated decision

Before we explain the obligations in more detail, it is important to make the following disclaimer. **These specific information obligations are (in principle) only applicable if 'automated decision making' is involved**, whether using profiling or not, and there are legal consequences for the data subject or if the data subject is otherwise *significantly* affected. We will further explain what this means in Art. 22 analysis.

For now, it is important to understand that the GDPR defines profiling in Article 4(4) as:

> any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements;

Profiling is composed of three elements (see Figure 5).

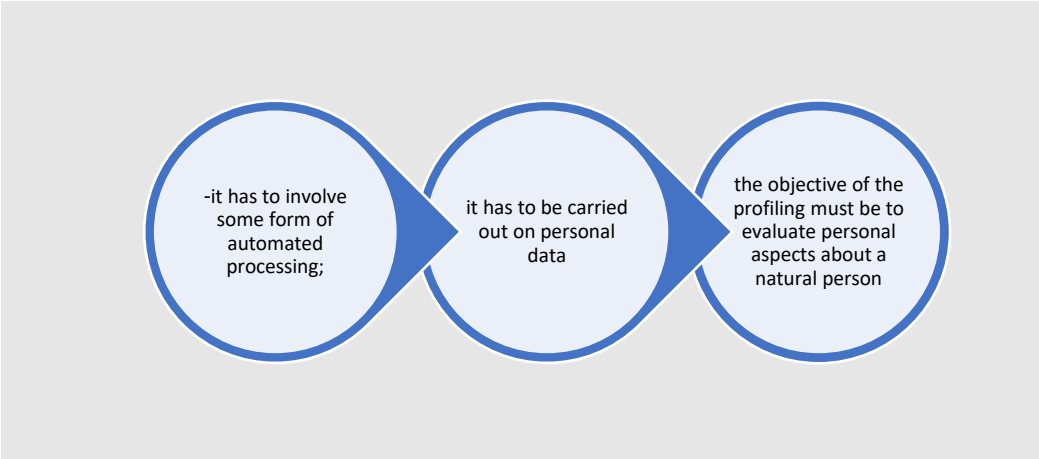

Figure 5: Elements of profiling

As explained by the WP29, *"profiling is a procedure which may involve a series of statistical deductions. It is often used to make predictions about people, using data from various sources*

requirements in Articles 13-14 and those in Articles 15. In the first case, the information must in principle be provided before or at the time of the processing of personal data. In the second case, the information will usually only be provided after the data subject requests it. The data subject can request this information at any time, including after the automated decision concerning her has been taken, with no deadline. As explained by the WP29 Guidelines "Article 15 implies a more general form of oversight, rather than a right to an explanation of a particular decision."

to infer something about an individual, based on the qualities of others who appear statistically similar" ('WP29 Guidelines on Automated Individual Decision-Making and Profiling', n.d.)

Profiling is therefore an automated processing of personal data for evaluating personal aspects, in particular to analyse or make predictions about individuals. A simple classification of individuals based on known characteristics such as age, gender for statistical purposes or to acquire an aggregated overview of its clients does not automatically lead to profiling. On the other hand, making predictions or drawing conclusions about individuals in order to e.g., provide a personalized news offer, is likely to qualify as profiling.

Automated decision-making has a different scope and can be based on any type of data. It may however partially overlap with or result from profiling. Solely automated decision-making is the ability to make decisions by technological means without human involvement.

❖ 1. OBLIGATION TO INFORM ABOUT THE EXISTENCE AND USE OF AUTOMATED (INDIVIDUAL) DECISION-MAKING AND PROFILING

Where controllers (or the processors they appoint) use automated decision making with personal data, they must inform about that fact the data subject. Data subjects therefore need to be informed when they interact directly with an AI system or when they communicate personal data to such systems. It is worth mentioning that the AI Act proposal (see section 4.1) imposes additional transparency obligations for systems that (i) interact with humans, (ii) are used to detect emotions or determine association with (social) categories based on biometric data, or (iii) generate or manipulate content ('deep fakes'), whether or not personal data is used.

❖ 2. OBLIGATION TO PROVIDE 'MEANINGFUL INFORMATION ON THE LOGIC INVOLVED'

The complexity of AI systems makes it challenging to explain (and understand) how an automated decision-making process or profiling works. The second transparency obligation is therefore about providing *a 'meaningful information about the logic involved'* in automated decision-making process.

○ What should the information be about?

According to the WP29 guidelines *"the controller should find simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision without necessarily always attempting a complex explanation of the algorithms used or disclosure of the full algorithm."* (WP29 2018). The guidelines do not elaborate on what "rationale" or "criteria" mean and how detailed should the information be. According to Bibal et al. "criteria" would mean providing all features with a non-zero coefficient in a linear model, or the features in a specific decision path of a decision tree, without necessarily providing the whole tree (Bibal et al. 2021). The issue lies in the size of such list of features used and the balance between accuracy and complexity of the model. As we will see later, the information should be 'meaningful'. The question raises how to efficiently present information to the data subjects?

"Rationale" could be interpreted as providing not only the features used in a decision, but also their combination used to make the particular prediction (Bibal et al. 2021). This can be done via transparent models such as decision trees or linear models, to create new ones (e.g., SLIM

(Ustun and Rudin 2016)) or to create ways to explain blackbox models (e.g., LIME, (Ribeiro, Singh, and Guestrin 2016)).

However, one should take a flexible approach in interpreting what "rationale" and "criteria" mean. Selbst and Powles rightly point out that *"one might think that meaningful information should include an explanation of the principal factors that led to a decision"*. However, such a rigid rule may prevent beneficial uses of more complex ML systems such as neural nets, even if they can be usefully explained another way (Selbst and Powles 2017).

- What does 'meaningful' mean?

The test for whether information is meaningful should be functional (Selbst and Powles 2017). In this context, such an explanation could have an instrumental value or an intrinsic one (Selbst and Powles 2017). The first interpretation seems to be favored by the WP29: The understanding is not the aim in itself. **The meaningful information is a mean to help a data subject *act* rather than merely *understand* the logic behind the decision-making process** (WP29 2018).

- 'Meaningful' to whom?

Finally, one may ask: meaningful to whom? Because Articles 13–15 of the GDPR all relate to the rights of the data subject, meaningful information should be interpreted in relation to the data subject concerned. That is, the information must be meaningful to an individual confronted with such decisions (Selbst and Powles 2017). According to the WP29, the information provided should be "sufficiently comprehensive for the data subject to understand the reasons for the decision" (WP29 2018).

To sum up, a data controller is not expected to provide a complex explanation of the AI system used, and definitely not to disclose an algorithm, the underlying source code or trade secrets (Knowledge Centre Data & Society 2021). This is not what the GDPR meaning of 'transparency' is about. Rather, it is important to inform a data subject concerned in an easily understandable, but useful and meaningful way about the underlying logic of the automated decision-making processes. Knowledge Centre Data & Society suggests that in order to comply with the transparency obligations, the following information can be communicated to the data subject (Table 3).

Table 3: Information to be communicated to the data subject

Information to be communicated to the data subject
The categories of data/information (and related attributes) that were or will be used in the (re)training, testing or operational use of the profiling or automated decision making systems. These include the personal data collected and how it is collected, the data quality or age of the data.
How the necessary measures were taken to ensure that the training and test data were (and still are) representative of the target group(s) for whom the aim is to make predictions or decisions.
Why these categories are considered relevant and their respective weightings.
How the model/profile used in the automated decision making process is constructed, including any relevant statistics used in the analysis.

Why the profile is relevant to the automated decision making process or what purpose is intended for.
How the profile is used to make a decision about the data subject and what criteria are used (e.g., the main methodological choices regarding, inter alia, the algorithms and/or model structure used, the way in which any parameters are determined and how they contribute to a decision).
The performance or accuracy of the underlying model, as tested on independent and representative test data.
To what extent human control and/or intervention is (possible) on the processing.

The authors conclude that *"this rather simple information will be more relevant for the data subject than the underlying mathematical mechanisms and will therefore contribute to the transparency of the processing."* (Knowledge Centre Data & Society 2021).

❖ 3. OBLIGATION TO INFORM ABOUT THE 'SIGNIFICANCE AND THE ENISAGED CONSEQUENCES' OF THIS PROCESSING FOR THE DATA SUBJECT

This obligation means that a data controller must provide the data subject information about the intended or future processing, and how the automated decision-making might affect the data subject. In order to make this information meaningful and understandable, real, tangible examples of the type of possible effects should be given (WP29 2018).

➢ "The right to explanation"

In recent years, the focus of many legal scholars has been on the meaning of explainability from the data protection law point of view. The core debate has primarily focused on whether or not the GDPR creates a *right* to explanation of automated decisions. In 2016, Goodman and Flaxman argued that the GDPR creates a 'right to an explanation' of algorithmic decision-making (Goodman and Flaxman 2016). That claim sparked a critical discussion. In 2017, Wachter et al. elaborated on the fact that a legally binding *right* to explanation, popularly imagined as a right to explanation of specific automated decisions of the type, does not exist in the GDPR (Wachter, Mittelstadt, and Floridi 2017). They argue that a non-existing 'right to explanation' in the GDPR should not be mistaken with other GDPR mechanisms: (i) information duties of data controllers (Articles 13–14); and (ii) the right to access to information (Article 15), and (iii) the right not to be subject to automated decision-making and safeguards enacted thereof (Article 22 and Recital 71). We explained the two first points above. We will now move to the third point.

To untangle what the 'explainability' in the context of the GDPR means, we will start with the basics.

According to Art. 22(1) of the GDPR:

> The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

This provision in itself creates a lot of interpretation questions. First, is it a 'decision' what an AI/ML system actually produces? Second, what does 'solely' mean? Third, what are legal or 'similarly significant' effects?

For now, it suffices to note that in short, 'solely' automated decision means there is no human involvement in the decision process. As provided by the WP29: *"an automated process produces what is in effect a recommendation concerning a data subject. If a human being reviews and takes account of other factors in making the final decision, that decision would not be 'based solely' on automated processing."* (WP29 2018). Importantly, a data controller cannot avoid the Article 22 provisions by fabricating human involvement. As an example, if someone routinely applies automatically generated profiles to individuals without any actual influence on the result, this would still be a decision based solely on automated processing. Otherwise, a narrow interpretation of 'human involvement' would open a loophole whereby any human involvement in a decision-making process could mean it is not 'automated decision-making'.

The GDPR does not define 'legal' or 'similarly significant' however, they should be understood as having serious impactful effects. Examples of 'legal effects' include automated decisions about an individual that result in, for example, cancellation of a contract, denial of a social benefit granted by law etc. Recital 71 of the GDPR provides the following typical examples of what 'similarly significantly affects him or her': automatic refusal of an online credit application or e-recruiting practices without any human intervention.

In short, for data processing to significantly affect someone the effects of the processing must be sufficiently great or important. The decision must have the potential to:

- significantly affect the circumstances, behaviour or choices of the individuals concerned;
- have a prolonged or permanent impact on the data subject; or
- at its most extreme, lead to the exclusion or discrimination of individuals (WP29 2018).

Much more can be said about the interpretation of this provision. However, this goes beyond the scope of this deliverable. More important for the discussion is the fact, that there are the following exceptions from this prohibition. A decision based solely on automated processing is allowed, where the decision is:

- necessary for the performance of or entering into a contract;
- authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subjects' rights and freedoms and legitimate interests; or
- based on the data subjects' explicit consent.

Note that automated decision-making that involves special categories of personal data is only allowed under the additional conditions (Article 22(4)).

To conclude this part, Article 22 provides that: (i) as a rule, there is a general prohibition on fully automated individual decision-making, including profiling that has a legal or similarly significant effect; (ii) there are exceptions to the rule. Moreover, (iii) where one of these exceptions applies,

there must be measures in place to safeguard the data subjects' rights and freedoms and legitimate interests safeguards, such as the right to obtain human intervention and the right to challenge the decision (Article 22(3)). We will describe them more in detail below.

- Establishing appropriate safeguards

As mentioned above, where automated decision-making meets a condition specified in Article 22(3)a (to enter or fulfil a contract) or Article 22(3)c (with explicit consent), data subjects are granted additional safeguards, including at least:

- the right to obtain human intervention on the part of the controller,
- to express his or her point of view,- and
- to contest the decision.

Again, the controller must provide a simple way for the data subject to exercise these rights. The data subject will only be able to challenge a decision or express their view, if they fully understand how it has been made and on what basis (WP29 2018).

Critically, a right to explanation is not mentioned in Article 22. In all of the GDPR, a right to explanation is only explicitly mentioned in Recital 71, which states that a person who has been subject to automated decision making

> "should be subject to suitable safeguards, which should include specific information to the data subject and the right (…) to obtain an explanation (…)"

The interpretation of Recital 71 has been the main bone of contention in the 'explainability' discussions. If legally binding, this provision would require an *ex-post* explanation of specific decisions. This is however, not the case. Recital 71 does not establish a legally binding right, because Recitals provide guidance on how to interpret the articles, but are not themselves legally binding. While they can help to explain the purpose and intent behind legally binding Articles, they themselves, do not impose any rule or obligation.

Finally, the question raises how to comply with all these information and transparency obligations in practice. Knowledge Centre Data & Society suggests the following actions the data controller can take (Table 4).

Table 4: How to comply with transparency obligations

How to comply with transparency obligations
Ensure that the organisation has a privacy statement that contains all the information required by the GDPR and is communicated to the data subjects at the appropriate time
Consider working with a layered privacy statement, especially if useful information relating to the underlying logic of the AI system needs to be provided
Consider using visual and interactive techniques to communicate this information to the data subjects in a clear and understandable way

Identify which processing operations using AI systems involve automated decision making and whether these processing operations entail legal or other significant consequences with respect to the data subjects
When developing AI systems, try to use an 'explainability by design' approach and strive for the most transparency design of AI systems possible
Inform data subjects as soon as they interact with an AI system that involves automated decision-making
Inform the data subject about the intended or expected consequences of the processing, using tangible examples

- Interim conclusion: reflections on transparency and 'explainability'

First, one cannot disagree with Hildebrandt that *"we should not mistake the legal obligation to justify actions or decisions for the right to explanation and/or information, even though they are clearly related. Explanation in itself does not imply justification, and justification does not always require an explanation of the underlying logic of the decision system."* (Hildebrandt 2019). It is important to underline, that transparency and 'explainability' are not *instead of*, but *rather next to* other data protection principles as explained in this deliverable. A decision of an automated system should be justifiable independently of how the system came to its conclusion.

Second, we need to be mindful that in many cases what the data subject wants is not an explanation—but rather for the disclosure, decision or action simply not to have occurred (Edwards and Veale 2017). Third, *"the law is restrictive, unclear, or even paradoxical concerning when any explanation-related right can be triggered"* (Edwards and Veale 2017). Indisputably, however, whether one uses the phrase 'right to explanation' or not, the data controllers still have to give certain information to the recipients of decisions including the meaningful information about the logic involved, as well as the envisaged consequences of such processing for the data subject (art. 13(2f) and 14 (2g) of the GDPR).

Importantly, the question raises how to reconcile legal interpretation of transparency obligations with technical capabilities of explaining AI models. As an example, Hamon et al. used a COVID-19 use case scenario to assess the feasibility of legal requirements on algorithmic explanations. They concluded that the use of complex deep learning models in AI applications, such as in COVID-19 detection, makes it hard to reconcile with the existing EU data protection law requirements, especially with regards to human legibility of explanations for non-expert data subjects (Hamon et al. 2021). In other words, the quality of possible explanations of the more advanced forms of decision-making, may not be found adequate under the GDPR. Similarly, Edwards and Veale note that the legal concept of explanations as "meaningful information about the logic of processing" may not be provided by the kind of machine learning "explanations" computer scientists have developed (Edwards and Veale 2017). They argue however that "subject-centric" explanations (SCEs) focusing on particular regions of a model around a query *"show promise for interactive exploration, as do explanation systems based on learning a model from outside rather than taking it apart in dodging developers' worries of intellectual property or trade secrets disclosure"* (Edwards and Veale 2017).

In an attempt to reconcile legal and technical perspectives, Wachter et al. argue for "counterfactual explanations". In their view, "*explanations of automated decisions need not hinge on the general public understanding how algorithmic systems function. Even though such interpretability is of great importance and should be pursued, explanations can, in principle, be offered without opening the black box."* (Wachter, Mittelstadt, and Floridi 2017). The "counterfactual explanations" aim to clarify for individuals targeted by automated decisions, amongst others, *"what would need to change in order to receive a desired result in the future, based on the current decision-making model*."(Wachter, Mittelstadt, and Floridi 2017).

3.1.2 Purpose limitation

Article 5(1)(b) of the GDPR provides that following the purpose limitation principle, personal data should only be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes. This principle is nevertheless nuanced as further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation'). The notion of purpose is very much linked to the legal basis as the data subject can consent to one or specific purposes for instance. The purpose is also seen as the guiding heading of the processing activity. It is crucial to understand that the purpose that must be defined by the data controller is not the same as defining the purpose by the data scientists. The first defines the *purpose of the controller*, the second defines a purpose *for the learning algorithm*.

The purpose limitation principle is composed of several elements. Firstly, the purpose must be **specified** purpose. Specified means a purpose sufficiently precise, hence the WP29 considered general statements such as for advertising purposes, for commercial purposes not meeting this requirement (WP29 2013). However, the context very much plays a role in the specification of the purpose as well as the data subject's expectations. Secondly, the purpose must be **explicit,** unambiguous and clearly expressed in an intelligible form for the audience targeted (WP29 2013). Thirdly, the purpose must be **legitimate,** meaning in respect in all laws applicable to the situation. Codes of conduct, contractual arrangement, all circumstances of the situation can also be taken into account, which is an interesting component considering the bloom of AI ethics codes of conduct (Biega and Finck 2021).

A lot of machine learning operations are based on repurposing data and hence raise several issues regarding the function creep. The function creep is referred as *"the expansion of the intended use of technology to a different use, bringing with it a series of unintended and uncontrolled consequences"* (Emanuilov et al., n.d.). In a data protection context, this would refer to the risk that the data are used for secondary purposes which are not compatible with the purposes for which the data were initially collected (Koops 2020; Kindt 2007; Wisman 2013). AI systems are programmed to perform a certain task, but the way this purpose is specified may trigger an optimisation process, a deviation from the designer's initial intention. When the model is applied to a different context compared to the one it was trained for, or the technology designed is used for a whole new purpose (Emanuilov et al., n.d.), it can impact the compliance with the purpose limitation principles and other fundamental rights.

- Principle

Sartor puts forward that *"the requirement of purpose limitation can be understood in a way that is compatible with AI and big data, through a flexible application of the idea of compatibility, which allows for the reuse of personal data when this is not incompatible with the purposes for which the data were originally collected."* (Sartor et al. 2020) The trigger is therefore the legitimate compatibility of the new purposes with the purpose used for the initial personal data collection.

- Repurposing for compatible use

The EDPB ancestor, the WP29, had adopted an opinion in 2013 on purpose limitation, which already clarified some aspects of personal data re-use (WP29 2013). The opinion established key factors to be considered during the purpose compatibility assessment. The assessment for compatible use is a case-by-case analysis.

Firstly, the relationship between the two purposes: the bigger is the distance between the two, the harder will it be to have a compatibility.

Secondly, the context in which the data have been collected and the reasonable expectations of the data subjects as to their further use must be considered. This includes the factual context of the processing such as the nature of the data subject/controller relationship, the contractual and legal obligations applicable. The more specific and restrictive the context of the collection, the more limitations there are likely to be on further use.

Thirdly, the nature of the data and the impact of the further processing on the data subjects. The aim is aligned to the GDPR's objective which is to protect individuals against the impact of improper or excessive use of their personal data. The nature of the data is assessed: the more sensitive the data is, the narrower the scope of compatible use will be. The consequences include potential future decisions and actions by third parties, the severity, the likelihood of the consequences on the individual's life. Could alternative or less intrusive measures be used to achieve the purposes?

Fourthly, the safeguards applied by the controller to ensure fair processing and to prevent any undue impact on the data subjects also matter. Preventive measures, compensation measures both at the technical and organisational level are considered.

Despite the relevance of these criteria, room for unclarity still exists when applied to AI applications (Sartor et al. 2020). The authors pointed out that the purpose limitation is often overlooked at the operation level and does not lead to concrete enforcement measures (Emanuilov et al., n.d.).

While including personal data in a training set, containing already a considerable amount of data, is not going to affect a specific person in particular; it nevertheless opens the door to data breach and misuses. Sartor argues that to avoid this, the personal data should be anonymised and deleted once the model is constructed. The compatibility requirement will be harder to meet when sensitive data are at stake and the additional safeguards to ensure a compensation

for the repurposing such as security measures, anonymisation and pseudonymisation are necessary to ensure the legitimate use. (Sartor et al. 2020)

In addition, even if the individual is not directly impacted by the inclusion of his personal data to the set, he is however directly affected as the personal data are effectively used in an algorithmic model. The data will be used to find common patterns with other individuals, the data subject may be categorised in a certain group and profiling may be conducted. When in presence of data sets used for profiling, the criteria of the compatibility test must be strictly applied (Sartor et al. 2020).

Others put forward that a more strict and in-depth analysis of purpose limitation is necessary at the design stage to avoid the function creep and to better understand the acceptable level of AI deployment (Emanuilov et al., n.d.). From a data subject's perspective, it is however not easy to verify if the data collected are processed in a way which corresponds to the purposes presented. Concerning AI media applications, Asia J. Biega and Michèle Finck reported that while Google and Netflix provide purposes and illustration for the use of personal data and these purposes, there are virtually no means of checking whether these verbal expressions correspond to what happens in practice (Biega and Finck 2021; see also Section 3.4 on data subjects' rights). Some authors indicate that explicitness could be achieved with purposes published in a machine-readable format (Koops 2011), listing browsing cookies explicitly in the privacy policy, developing purpose standards (Fouad et al. 2020). Others put forward that using service improvement could be considered as an objective criterion for purpose formulation in an AI system context but only if the purposes were legitimate, explicit and specific enough (Biega and Finck 2021).

- GDPR explicit exceptions to re-use compatibility test

Article 5(1)(b) of the GDPR provides that personal data can also be further processed for archiving purposes in the public interest, scientific or historical **research** purposes or statistical purposes in accordance with the safeguards listed in Article 89 GDPR. Recital 159 adds that the processing of personal data for scientific research purposes should be interpreted in a broad manner, including for example technological development and demonstration, fundamental research, applied research and privately funded research. It is not crystal-clear what research encompasses and some wonder if research teams from private companies or researchers working part-time for companies could be falling in the scope of this exemption (Biega and Finck 2021). They point the non-incentive that this inclusion could create for purpose limitation principles and a potential way for big players to circumvent the safeguards established by the GDPR.

According to article 5 (1)(b) and Rec. 50 of the GDPR, reuse for **statistical purposes** is assumed to be compatible with purpose limitation unless this involves unacceptable risks for data subjects. Statistical purposes are defined by Rec. 162 as a processing necessary for statistical surveys or for the production of statistical results having as output aggregated data and the results must not be used in support of measures or decisions regarding any natural person. Hence, authors argue that some AI systems processing operations could be categorised as statistical purposes if they respect the above conditions (Biega and Finck 2021). The authors

provide an interesting illustration: “a search engine might train a non-personalized ranker that preselects webpages as a response to a query, and then use an individual’s personal data to re-rank the webpages in the preselected set. In a scenario like this, the training of the aggregate model might be considered a form of statistical analysis (and thus not subject to data minimisation), while applying the model in conjunction with an individual’s data will not (as it produces individual results).” The GDPR leaves the door open to the Member States to adopt more specific rules on statistical purposes.

- GDPR solution to incompatible purposes

Recital 50 of the GDPR provides that where the data subject has given consent or the processing is based on Union or Member State law which constitutes a necessary and proportionate measure in a democratic society to safeguard, in particular, important objectives of general public interest, the controller should be allowed to further process the personal data irrespective of the compatibility of the purposes. The controller must ensure the data subject’s right to object. This solution is apparently widely used by data controllers to legitimise further processing especially for direct marketing, behavioural or location-based advertisement, data-brokering, or tracking-based digital market research (Biega and Finck 2021; WP29 2013). However, the reliance on consent has been heavily criticized by privacy, data protection advocates and consumer rights associations. They put in question if the consent is truly “informed” and “freely given” component given the power imbalance between data subjects and controllers, especially in the absence of technical tools to efficiently implement the consent revocation consequences (Noyb 2018, Biega and Finck 2021).

3.1.3 Data minimisation

Article 5 (1)(c) of the GDPR provides that personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (data minimisation). The data minimisation principle is a follow-up to the principle of purpose limitation analysed above. The data processed should be strictly necessary to achieve the purposes outlined by the controllers and is strongly connected to the idea of proportionality.

The GDPR requires to have data relevant, adequate and necessary for the processing. Firstly, **relevant** data helps preventing “accumulation of data for the sake of gathering data or for undisclosed ends”. This could be to use it for future new purposes or for re-sell (Biega and Finck 2021). Some have argued that the amount of data collected is justified by the need to provide accuracy to the AI system outputs and that the quantity is not much the issue but rather the use which is being made thereof (van Hoboken 2016). Others expressed how the principle of data minimisation is preventing AI systems to fulfil their potential, limiting innovation and sacrificing social benefits (MacCarthy 2018). It is true that AI systems are pushing the limits for what is relevant. The more data are collected, the more meaningful the combination and connections become, this deepens the analysis and AI potential to some extent (Biega and Finck 2021).

Secondly, **adequate** data enable to have a fair, transparent and accurate model. In some cases, adequacy will limit or increase the collection of data to ensure the implementation of these characteristics. This has proven true especially for underrepresented demographic groups.

Thirdly, the **necessity** requirement requires the controller to identify the minimum amount of personal data needed to fulfil a purpose. For instance, if less data or anonymous data could achieve the identified purposes, the personal data would be deemed unnecessary.

Data can be added in course of the processing or retained for longer, if they provide a benefit for the purposes of the processing, while balancing the risks for data subjects and including security and organisational measures (Sartor et al. 2020).

- What should be minimized?

From the more traditional perspective, which expects that the quantity of personal data be minimised, we see different trends emerging, which could match the AI development and need for data. Indeed, authors allege that not only data quantity could be reduced but also characteristics associated with data including the identifiability of the data subjects or the sensitiveness of the data used (Biega and Finck 2021). Sartor puts forward that in order to be compatible with AI systems development, the data minimisation principle could be interpreted as not only reducing the amount of data needed but also reducing their "personality" level (Sartor et al. 2020). This interpretation uses several measures including pseudonymization, anonymization and considers the re-identification likelihood and facility. His interpretation argues that re-identification of personal data after pseudonymisation or anonymisation shall be prohibited and that in case where a data can be re-identified to a data subject, it shall be considered as a new personal data (triggering the GDPR requirements) unless all conditions for the lawful collection of personal data are met and the processing is compatible with the purposes of the initial collection.

In recent years, a new research trend emerged on data minimization, which argues that blindly adding vast amount of data leads to diminishing returns in model performance (Hestness et al. 2017; Sun et al. 2017; Shanmugam et al. 2021). Some algorithmic techniques now exist and permit to reduce the amount of data processed while improving the accuracy such as outlier detection, feature selection, learning framework with data collection stopping points *(*Biega and Finck 2021*;* Shanmugam et al. 2021). Other systems enable to use anonymisation techniques to suppress and generalise input features in classification (Goldsteen et al. 2021).

However, the GDPR principles are equal and must all be fulfilled for a compliant processing activity. This complexifies the implementation of data minimisation and even if techniques can be found to solve the tension between minimisation and accuracy, there is another tension with fairness which will be harder to solve (Shanmugam et al. 2021). Authors point out that data minimisation could harm the marginalized population and therefore requires specific safeguards (Shanmugam et al. 2021; Wen et al. 2018).

In practice, this principle seems hard to implement as a recent study showed how there are no common practices among software practitioners due to a lack of practical guidelines (Senarath and Arachchilage 2018) and there is apparently no such study on AI systems developers (Biega and Finck 2021). Furthermore, the comprehensive analysis of Biega and Finck points out that there is currently no method to identify which data improves results in an AI system. In addition, another component to keep in mind is that minimisation of data for a single user will also

influence the performance of the system for other users. Lastly, the two authors also point out that it is difficult to infer user intent from their behaviour and an inaccurate detection of user personal purposes might lead to both under- and over-minimisation of data, depending on the context as an AI system could minimize some data in relation to the user intent (Biega and Finck 2021).

Despite their complex practical implementation, purpose limitation and data minimisation principles help reduce the noise in the data and are therefore not only data protection safeguards but also successful AI systems safeguards (Biega and Finck 2021). Nevertheless, the authors point out the following difficulties in implementing the principles in practice (Figure 6).

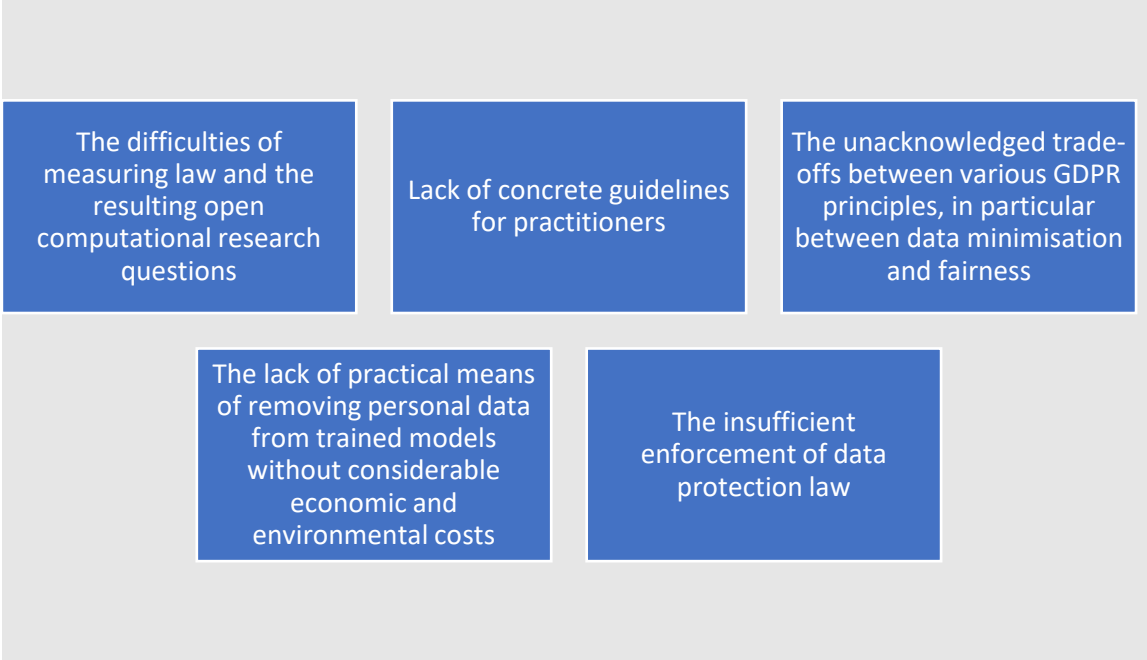

Figure 6: The practical difficulties of implementing data minimisation and purpose limitation

Biega and Finck propose the following solutions to mitigate these challenges (Biega and Finck 2021) (Table 5).

Table 5: Possible solutions to mitigate data minimisation and purpose limitation implementation challenges

Possible solutions to mitigate data minimisation and purpose limitation implementation challenges
Stimulate a business culture on data minimisation and purpose limitation

Develop tools and techniques to enforce these principles: features selection, data influence estimation, data valuation, active learning for prioritisation purposes
Ensure that the principles are enforced from the design phase and all along the AI system
Support research on mathematical interpretations of the principles and machine learning models for automatic compliance and understanding the effect of data subjects' rights enforcement on the overall fairness
Develop standards for data processing purposes
Establish framework for removing data from existing models
Set up auditing methods

3.1.4 Accuracy

As reported by several authors, the accuracy principle has up until now received only limited attention in data protection studies and remains an unexplored legal question not challenged by courts (Biasin 2021; Dimitrova 2021). However, it recently grasped attention because of the bias and errors risks linked to personal data uses by AI systems (Biasin 2021).

Article 5(1)d of the GDPR requires controllers to ensure that the personal data are 'accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay'. This principle has pertaining data subject right in the right to rectification and the right to restriction of processing (see section 3.4). The accuracy principle can be seen as the data subject's right enabler and is very much dependent on the purpose and context at stake (Dimitrova 2021). There is also a trend in replacing data accuracy by data quality which seems to, besides a denomination change, constitute a multidimensional concept going beyond data accuracy (Dimitrova 2021). Dimitrova also argues that following this observation, data accuracy and the right to rectification should be broadly interpreted.

The accuracy principle would apply to personal data used as input to AI systems. For instance, inaccurate data can influence the decision taken by AI systems about the data subjects and expose him or her to harm (Sartor et al. 2020). In addition, in a big data environment, the enforcement of the accuracy principle remains unclear as AI systems are made for establishing correlation, finding patterns, inferring predictions and probabilities and hence deducting information about data subjects that could turn inaccurate (Naudts et al., n.d.). However, authors also pointed that accuracy could encompass not only the input data but also the design of the algorithm using personal data (Malgieri and Comandé 2017). This could lead to incorrect assessment of the data, the associated data analysis and the ensuing incorrect results (Hoeren 2017). The Court of Justice of the European Union even pointed out that the accuracy requirements in light of privacy fundamental rights should ensure that the criteria and models are 'specific and reliable' to fulfil the processing purposes (Opinion 1/15 of the Court (Grand Chamber) 2017). This comes with a clear conceptualisation of the purposes, to ensure a reliable interpretation of the results (Dimitrova 2021).

Therefore, several elements could be taken into account to identify the inaccuracy source: inference resulting from the processing (such as a poor statistical method), inference from wrong input data, and lastly correct inferences but inaccurately predicting the outputs (Hallinan and Borgesius 2020).

The level of accuracy needs to be in line with the processing purpose(s). A delicate balance needs to be struck as without enough accurate data, the purposes of the processing might not be achieved. Paradoxically, as Chen points out, a high degree of accuracy may bring its shortcomings such as new forms of discrimination and the loss of individual manoeuvre space (scoring systems) (Chen 2018).

The WP29 released guidelines outlining that errors or bias in collected or shared data or an error or bias in the automated decision-making process can result in incorrect classifications and assessments based on imprecise projections that impact negatively on individuals (WP29 2018). The same guidelines provide that regular checks need to be conducted to spot and solve bias in the data but also that auditing systems should be put in place to ensure accuracy and relevance of automated decision-making. It is true that the risks for inaccuracy and discrimination were already spotted by the WP29 in 2013 among the risks and challenges posed by big data to the right to personal data and privacy (WP29 2013).

A distinction is being operated by Sartor between personal data which are a part of a training set and personal data which are a part of a profiling algorithm (Sartor et al. 2020). He explains that it is very tempting once personal data is part of a training set to use the same data for establishing individualized preferences. He further expresses that anonymisation, pseudonymization and security measures will be key to addressing the risks stemming from this temptation.

Preventive measures could also help ensure accuracy through firstly conducting a data protection impact assessment and then ensuring close monitoring of substantial/procedural safeguards when it comes to automated decision-making processes (Naudts et al., n.d.). Involving data subjects more actively in data management and rectification has also been advanced for a long time as a way to ensure accuracy (Karst 1966).

- Interim conclusion: reflections on purpose limitation, data minimisation and accuracy principles

There is a growing tension between the full deployment of AI and big data and these three GDPR principles. Because AI systems need a considerable amount of data to find correlations and establish predictions, the respect of these principles seems difficult, and sometimes, impossible to reach. This is especially relevant as AI media applications are widely based on algorithmic profiling, personalisation, and decision-making. They do not only collect 'traditional' personal data such as name, location, gender but also more insidiously *"behavioral interaction logs (such as search queries, product ratings, browsing history, or clicks"* (Biega and Finck 2021).

From the outset, the purpose limitation and data minimisation seem hardly compatible with the essence of AI systems which re-use personal data for new purposes repeatedly in order to reach their potential (Mayer-Schönberger 2016). The opportunity and relevance of these two

principles have been heavily discussed in practice. Some think that considering AI, big data, and Internet of Things development, the two principles will need to be abandoned (Moerel and Prins 2016).

Despite the visible conflicts between embracing AI systems' potential and data protection, authors found ways to interpret the GDPR principles in line with the AI development (Sartor et al. 2020) and others observed that systems could technically use much less data than they currently do (Biega and Finck 2021).

3.1.5 Storage limitation

Art. 5(1)(e) of the GDPR provides that *"personal data shall… 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."* In other words, the principle of storage limitation means that personal data must be deleted or made anonymous as soon as they are no longer necessary for the purposes for which they were collected. When data is no longer needed, it should either be erased or anonymised. The GDPR itself does not impose specific storage periods for different types of data. It is up to the data controller to determine this and it will depend on how long the data is needed for the specific processing operation. Some national law provisions, may, however, determine the maximum storage period.

Additionally, in order to ensure appropriate implementation of time limits, along with Art. 25 of the GDPR, it is mandated that controllers implement appropriate technical and organisational measures for ensuring, by default, the legitimate period of storage of personal data is respected, e.g., in the form of expiry dates for each set of data. The general rule is therefore that personal data may not be stored indefinitely, nor may they be stored solely because they might be 'useful' in the future. The EPRS and others provide that there is *undoubtedly* a tension between the AI-based processing of large sets of personal data and the principle of storage limitation (Sartor et al. 2020). However, to mitigate this tension, the GDPR provides an exception for archiving, research or statistical purposes, which is of interest for AI researchers.

3.1.6 Integrity and confidentiality (security)

Art.5 (1)(f) of the GDPR defines the principle of 'integrity and confidentiality', which is a crucial requirement of security in data processing. The provision sets forth:

> *"Personal data shall… be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction and damage, using appropriate technical or organisational measures."*

Personal data can only be properly protected, if measures are taken to ensure their integrity and confidentiality. On the one hand, this refers to technical measures such as implementing encryption, a firewall or password control. On the other hand, organisational measures are also required, such as imposing certain obligations on staff and subcontractors.

These measures are intended to prevent the personal data being accidentally or unlawfully shared with or exposed to third parties or persons who should not have access to them (whether in bad faith or not), lost, destroyed or changed. These measures need to take into account the state of the art, the costs of implementation, the context and the risks for the individuals whose data are being processed. Security is therefore a dynamic obligation that is risk- and context-based and that can also evolve within the same organisation (Knowledge Centre Data & Society 2021).

3.1.7 Accountability principle

The accountability principle is one of the basic principles of data processing. It requires controllers and processors to be able to demonstrate that they have taken steps to comply with the obligations under the GDPR.

Despite being overly-used in contemporary scholarly literature and legal and policy instruments, what 'accountability' exactly means in practice is complex. WP29 Opinion on Accountability sheds some light of the meaning by stating that within this context, it means showing how responsibility is exercised, demonstrated, and made verifiable (WP29 2010). In other words, responsibility needs to be demonstrated as working efficiently in practice to be able to develop sufficient trust.

- What does the accountability principle entail?

Recital 74, read with Art. 5(2) of the GDPR, explains the main elements of the principle (Figure 7).

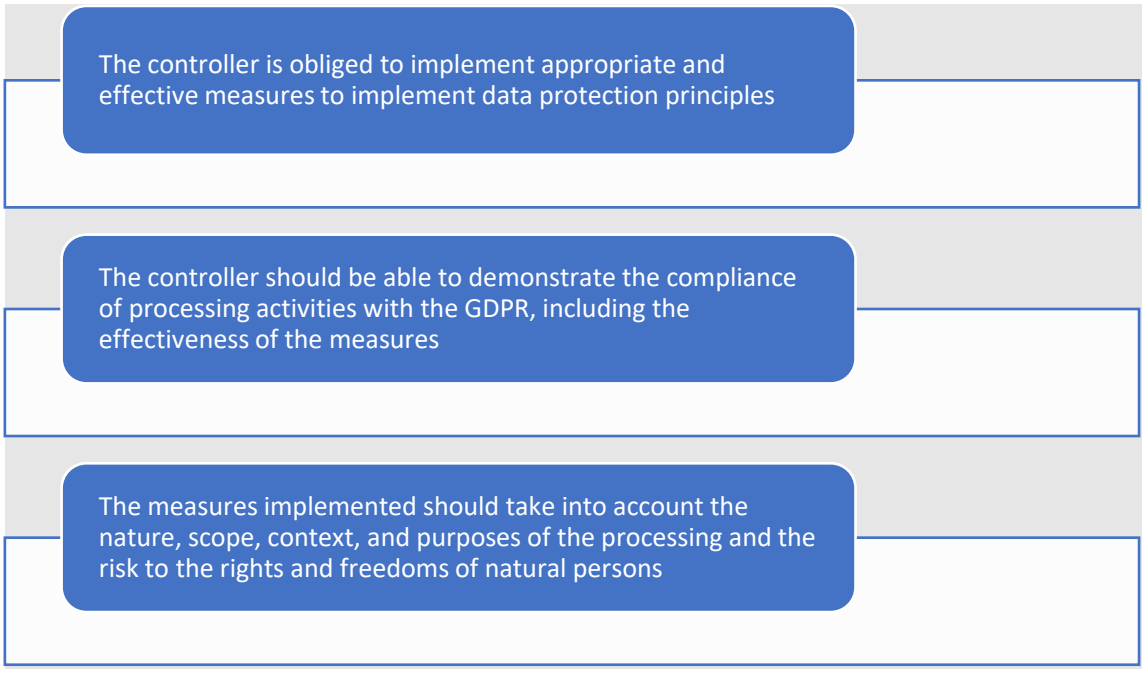

Figure 7: The main elements of the accountability principle

➢ How to implement accountability in practice?

In 2010 already, the WP29 favoured the introduction of accountability and stated that the expected results of accountability mechanisms would *"include the implementation of internal measures and procedures putting into effect existing data protection principles, ensuring their effectiveness and the obligation to prove this should data protection authorities request it"* (WP29 2010).

In the GDPR, many of these 'accountability' measures have been adopted. According to the EDPS, accountability in personal data processing involves transparent internal policies, training employees, responsibility at the highest level for the monitoring of the implementation, assessment and demonstration to external parties of the implementation's quality and procedure for redressing poor compliance and data breaches (EDPS 2016). The EPDS considers other examples of accountability obligations such as data processing documentation, the need to install data security measures, the requirement to make a data protection impact assessment and data protection by design and default.

Data protection by design and default are separate, yet related concepts. They constitute an integral element of being accountable, requiring embedding data protection into everything you do, throughout all your processing operations.

Article 25(1) specifies the requirements for data protection by design:

> 'Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.'

Data protection by design is ultimately an approach that ensures you consider privacy and data protection issues at the design phase of any system, service, product or process and then throughout their lifecycle.

Article 25(2) specifies the requirements for data protection by default:

> 'The controller shall implement appropriate technical and organisational 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.'

Data protection by default requires you to ensure that you only process the data that is necessary to achieve your specific purpose. It links to the fundamental data protection principles of data minimisation and purpose limitation.

Additionally, Article 25(3) states that another way of demonstrating compliance is to adhere to an approved certification under Article 42 of the GDPR.

3.1.8 Interim conclusion

Section 3.1 provided an analysis of the various GDPR principles and how they are interpreted in an AI context. The section also highlighted where clarifications are lacking and how gaps need to be filled to ensure a compliance between the GDPR and personal data processing for AI systems. The GDPR principles constitute a key guiding framework for personal data processing. However, one cannot forget, that the GDPR introduces many more legal obligations on data controllers and processors. This deliverable does not aim to explain them all. However, in the following section, we will focus on another important aspect for AI system compliance with data protection framework, namely the data subject's rights.

3.2 Data subject rights in the context of AI

GDPR's provisions impose obligations on controllers and processors and inevitably create a data subject's right dimension as the first ones process the latter's personal data. However, in addition to this aspect, the GDPR also created specific and stronger data subjects' rights contained in Chapter 3 of the GDPR. These rights will be detailed in the section below in light of AI systems considerations.

The GDPR does not specify under which format must data subjects lodge their request, nor does it provide any further information on the topic. It is advised for practitioners using AI systems processing personal data to outline procedures on enforcement request and process (Knowledge Centre Data & Society 2021). The deadline to reply to a right's request is important to keep in mind and controllers must reply to such request within a month except; where the request proves to be complex, an extra two months can be granted and notified to the data subject. If the request is refused, the reasons must be communicated to the data subject. All data subjects' rights requests must be addressed free of charge except when manifestly unfounded or excessive. 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.

➢ Right to be informed

Article 13 and 14 of the GDPR set a right to be informed (see Section 3.1.1.3). In line with GDPR principles and underlying considerations, controllers must ensure they explain clearly and simply to individuals how the profiling or automated decision-making process works. In an AI context, complying with this right appears challenging (see Section 3.3).

➢ Right not to be subject to a decision based solely on automated processing

As already explained above in Section 3.1.1.3, article 22 of the GDPR sets up a general prohibition and exception regime regarding automated decision-making, including profiling, which produces legal effects concerning the data subject or similarly affects him or her. If one of the exceptions of Article 22(3) (a) or (c) apply, the data controller must implement suitable measures to safeguard the data subjects' rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

➢ The so-called right to explanation

The 'right to explanation', mentioned in Recital 71 is one of the most debated concepts of the GDPR. See Section 3.1.1.3 for more information.

➢ Right of access

The right of access is of particular importance as it enables the data subjects to exercise the other rights provided for by data protection legislation as it enables transparency and accountability (EDPS 2020). Article 15 includes three different components of this right (Figure 8) (EDPB 2022).

Confirmation	Information	Copy
•A right to obtain a **confirmation** as to whether data about the person is being processed or not.	•A right to **obtain information** of any personal data used for profiling, the purpose, duration of the processing and the categories of data used to construct a profile, including: •the existence of automated decision making, including profiling; •meaningful information about the logic involved; and •the significance and envisaged consequences of such processing for the data subject.	•A data controller's duty to **make available a copy of all personal data processed,** even the data inferred from other data •A data controller should ensure a remote access to a secure system which would provide the data subject with direct access to his or her personal data. (Recital 63)

Figure 8: The components of the right of access

However, Recital 63 also nuances these obligations by indicating that the right of access *"should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software".* The copyright protecting a software is an example of this exception but the result of those considerations should not be a refusal to provide all information to the data subject.

The EDPB recently released comprehensive guidelines on the right to access and a consultation on the document is currently running before adopting a finalized version of the guidelines (EDPB 2022). One of the issues which raises is to what data should the data subject have an access to. The guidelines indicate that data inferred from other data, including algorithmic results and results of a personalisation or recommendation process, should be part of the personal data scope. They also included observed data and derived data from other data.

On how to provide access to personal data, the EDPB guidelines clarified that *"unless explicitly stated otherwise, the request should be understood as referring to all personal data concerning the data subject and the controller may ask the data subject to specify the request if they process a large amount of data"* (EDPB 2022). The information must be communicated in a clear, concise and audience adapted way. When considerable amount of data are processed, a layered approach could be undertaken only if it provides an added value for the data subject.

- Right to rectification

Article 16 of the GDPR provides that the data subject shall have the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her.

As often in AI context, the systems and the algorithms are designed to establish predictions, find patterns and correlations. This increases the risk of inaccurate, irrelevant, and out of context data used to fuel the systems and take decision impacting the data subject (WP29 2018). This is why, the data controller must ensure correctness and update data used for training an AI system (Knowledge Centre Data & Society, 2021).

The rights to rectification and erasure apply to both the '**input personal data**' (the personal data used to create the profile) and the '**output data**' (the profile itself or 'score' assigned to the person) (WP29 2018).

The right to rectification has the potential to be a powerful tool for data subjects to "to rectify not only factual mistakes, but also possibly profiling, risk assessments and data presentation problems" (Dimitrova 2021). However, his practical implementation needs to be refined by further guidelines or case law.

In the context of training data, Binns points out that individual inaccuracies are less likely to have any direct effect on an individual data subject (Binns 2019). In his view, *"requests for rectification of model outputs (or the personal data inputs on which they are based) are therefore more likely to be made, and should be treated with a higher priority, than requests for rectification of training data."*

➢ Right to erasure or also known as 'the right to be forgotten'

Article 17 of the GDPR provides that the data subject shall have the right to obtain the erasure without undue delay of his or her personal data if one of the listed grounds applies. An individual has the right to have their personal data erased if (Figure 9):

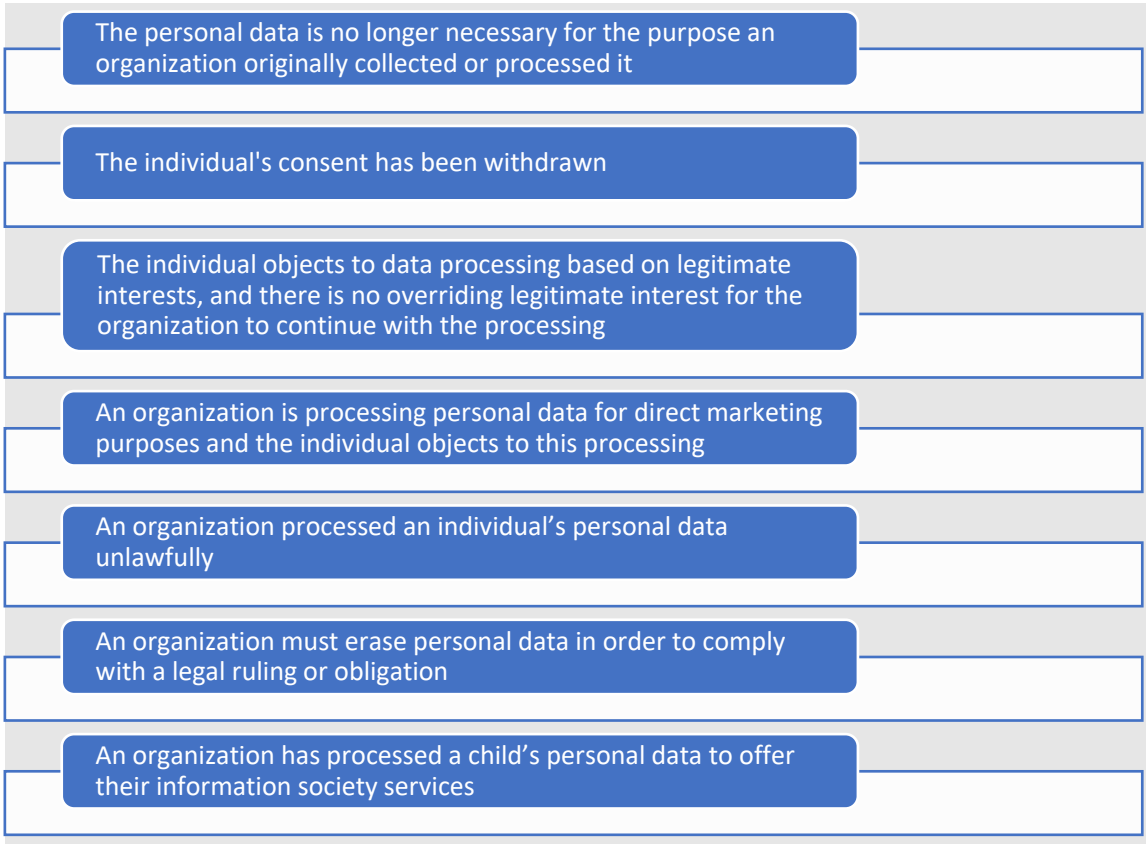

Figure 9: Grounds for data erasure right

In an AI system, context clarification could be brought as it is not clear what falls in the scope of personal data: should all data inferred also be targeted by the erasure request? Could this be suggested by the latest EDPB guidelines on data access. The guidelines include specifically inferred data based on the personal data processed in the scope of the obligation (EDPB 2022). Sartor suggests that only inferred data can be included in the scope of erasure right and not inferred group data (such as a trained algorithmic model) (Sartor et al. 2020).

The obligation is nuanced as there is room for exceptions. The list of exceptions contains freedom of expression, public security, legal claims reasons but also research or statistical purposes. Indeed, Article 17(3)(d) GDPR allows researchers to ignore an erasure request where it would render impossible or seriously impair the processing of personal data for scientific purposes. Thanks to this exception, personal data could be retained when the AI system would need re-training, quality search or evaluation processes. As provided by Sartor, "this limitation

would probably find limited application to big data, since the exclusion of a single record from the processing would likely have little impact on the system's training or, at any rate, on the definition of its algorithmic model" (Sartor et al. 2020).

In the context of training data, Binns points out that if the training data is no longer needed because the ML model has already been trained, the organisation must fulfil the erasure request. However, "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." (Binns 2019).

- Right to restrict processing

Article 18 of the GDPR provides that the data subject shall have the right to restrict processing of personal data when one of the conditions apply. The conditions include situations when the accuracy is contested, when there is unlawful processing, when the data subjects chose to opt for a restriction of use rather than an erasure request or when the personal data are no longer for the purposes of the processing. The right to restrict is immediate but time-limited and according to Veale and other authors “while it could in theory be used quite disruptively, is generally considered a lesser right to Article 21 objection” (Veale, Binns, and Edwards 2018).

The guidelines on automated Individual Decision-Making and Profiling provide that the right to restrict processing applies to any stage of the profiling process (WP29 2013). The right to restriction could be organised around opting in and opting out for personal data processing in an AI system context. Interestingly, as spotted by Fjeld et al. (2020), the High-Level Expert group guidelines on trustworthy AI had initially placed a positive obligation on data controllers to “systematically” offer an “express opt-out”. The final version only provides opt-out in case of citizen scoring technologies and when necessary to ensure compliance with fundamental rights (High-Level Expert Group on Artificial Intelligence 2019). In practice, it is not clear how should practitioners ensure compliance with this principle when operating an AI system.

- Right to data portability

Article 20 of the GDPR provides that the data subject has the *"right to receive the personal data concerning him or her, which he or she has provided to a controller in a structured, commonly used and machine-readable format' and 'to transfer the data to other controller".* Their right is only available when the processing is carried out by automated processes and is based on consent, which makes this right having a limited scope. Similarly to other rights, the scope of the right to data portability needs to be clarified when it comes to AI. In particular, the question raises if the right to data portability relates to all data related to the data subject or only the data that the data subject has actively provided (Sartor et al. 2020). Researchers observed that the following categories could fall under the category data provided by the person: (i) data actively and knowingly provided by the data subject; (ii) the observed data provided by the individual through the use of a service or device (search history, internet traffic, behaviour on a website) (Knowledge Centre Data & Society, 2021). The Centre argues that this concept does not include the data that the controller derives and deduces on the basis of the provided data.

Importantly, the data controller is not allowed to use the transmitted third party data to serve its own interests, for instance for proposing marketing products or for enriching the profile of a data subject. It shall also make sure to transmit personal data in a form that does not release information covered by trade secrets or intellectual property rights. As provided by Graef, how to precisely strike the balance between different interests and technicalities when exercising the right to data portability, requires further guidelines and case law. Despite the (non-legally binding) guidance from the WP29 on data portability, the concrete application of the right to data portability in practice still raises issues (Graef 2020).

Moreover, Kuebler-Wachendorff et al. point out that clear specification about the required *"structured, commonly used and machine-readable"* data format is lacking (Kuebler-Wachendorff et al. 2021) . Some authors argue that at present, only eight data formats (e.g., CSV, JSON, XML) are deemed fully compliant with the GDPR's requirements (Wong and Henderson 2019). Moreover, it is also unclear how a direct transfer between services should work from a technical perspective, especially due to lack of standardization, compatibility and interoperability between data formats.

- Right to object

Article 21 of the GDPR provides that a data subject can also refuse types of processing which they did not consent to, including processing based on 'legitimate interest' or a 'public task'. The processing can be objected as long as the data controller cannot demonstrate a 'compelling' legitimate interest. The bar seems high to meet unless some social benefit is present and not only economic interests of the controller (Veale, Binns, and Edwards 2018).

Article 21 of the GDPR provides that information on the right to object must be explicitly brought to the attention of the data subject in a clear and in a separate manner to the other information communicated according to the GDPR. Once exercised, the controller shall no longer process the personal data unless the controller demonstrates compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject or for the establishment, exercise or defence of legal claims. The GDPR does not provide further information on the compelling legitimate grounds, but the guidelines indicate that the controller may refuse to grant the right to object if he demonstrates that the processing is beneficial for society at large (or the wider community) and not just for the business interests of the controller. To demonstrate this compelling legitimate interest, the importance and the impact of the activity should be outlined in respect of the data minimisation and proportionality principles. The burden of proof lies on the shoulders of the controller using the exception.

Interestingly, Article 21(2) of the GDPR provides for what is being called by the EDPB an "unconditional" right to object to the processing of their personal data for direct marketing purposes, including profiling to the extent that it is related to such direct marketing (WP29 2018). The controller cannot argue or discuss the object's request, and shall free of charge access the objection request that the data subject can address at any time.

The right to object also applies to processing for scientific or historical research purposes and for statistical purposes. In such cases, the objection concerns the inclusion of the data subject's information in the input data for the processing at stake (as the result of research and statistics cannot consist in personal data)" (Sartor et al. 2020)).

3.3 Challenges to comply with data subject rights in big datasets

It is not always an easy task for data subjects to exercise their rights nor is it for controllers, including AI researchers, to comply with data subject requests. AI researchers or developers are typically looking for vast datasets in order to produce reliable and accurate outputs. These datasets often contain personal data and sometimes even sensitive personal data. The following datasets were and are used to train AI systems in order to get quality outputs from the system. For instance, for training AI models, data coming from Common Crawl are used for training large language models, ImageNet for object recognition or MS COCO for computer vision tasks. However, these datasets have come under close scrutiny after researchers and journalists flagged issues such as the lack of legal basis to process personal data contained therein, quality and representation issues, and, importantly for this section, issues related to data subjects' rights. Below, we present a selection of challenges encountered for complying or enforcing with data subjects' rights.

- Complexities related to the different stages of AI system processing

The data subject can exercise his or her rights at different stages of the lifecycle of an AI system processing personal data, namely the training, the output and the model stage (Knowledge Centre Data & Society, 2021).

The first potential challenge for fulfilling individuals' rights is the difficulty involved in identifying the individuals the datasets relates to. Indeed, singling out data belonging to a specific individual is not easy, if the data is part of a training system. However, this complexity does not make personal data any less personal. Therefore, the personal data located in a training system must be taken into account when a data subject wishes to exercise one of his/her rights (Knowledge Centre Data & Society, 2021).

As explained by the Knowledge Centre Data & Society, a model may sometimes "contain a set of individual examples that are part of the internal logic. This is done so that the AI system can distinguish with or between new examples during operationalisation". Therefore, even if only a small portion of the model contains personal data, there is still a chance that the data subject will wish to exercise their rights and the controller will have to comply with such a request. It is advised to think about this likelihood while designing the model in order to set up an easy retrieval process in order to comply in due time with the data subject request (Knowledge Centre Data & Society, 2021).

These requests can have a major impact on the AI system itself: from small adaptations and changes to the model to retraining the system or even destroying the model if the personal data cannot be separate from it, which may be challenging in practice (Knowledge Centre Data & Society, 2021).

- Transparency and right to information key for exercising the other rights

One major challenge relates to the right to information detailed in the section 3.1.1.3 above. Indeed, in the case of re-use of personal data contained in publicly available datasets, the personal data has not been obtained from the data subject itself. However, the data controller must nevertheless provide the data subject with information "within a reasonable period after obtaining the personal data, but at the latest within one month". The right to be informed is a prerequisite for a data subject to enforce his or her other rights. Otherwise, it is impossible to exercise data subject rights without being first aware of a data being (re)used.

Furthermore, this obligation is nuanced by exceptions, including where the data subjects already received the information, or when the data must remain confidential for professional secrecy. Article 14(5)(b) of the GDPR also provides an exemption for scientific research. The information obligation falls when it proves impossible to achieve or requires a disproportionate effort or risks seriously compromising the achievement of the research. Even then, however, the controller should, "take appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available" (Article 14(5) of the GDPR).

- Uncertainties regarding the application of data subjects' rights

As demonstrated in the section above, a lot of uncertainties remain about the scope of the data subject's rights when it comes to AI. It seems from the analysis conducted that clarification is necessary in order to, on the one hand, empower individuals to know the scope of their rights in an AI context and, on the other hand, to ensure AI practitioners legal certainty for their activities.

- Unfriendly AI system interface for rights enforcement

Even if clarification is being brought about the scope of the rights and data subjects become more aware of their rights, this must go along with friendly design. A consultation conducted for the ATAP project showed that if the design of the interface makes it hard for data subjects to exercise their rights, they show a tendency to express a "why even try?" approach (Lambrecht et al., n.d.).

- Lack of enforcement leads to trade-offs

The GDPR enforcement is a complex topic, but many describe it as suffering from enforcement lack. This creates a lack of incentive for controllers who *"rationally make a trade-off between the economic benefits of unconstrained usage of personal data and the potential yet very unlikely economic cost of data protection enforcement"* (Biega and Finck 2021).

4. Upcoming European legislation relevant to the provisions of the GDPR

This section aims to reflect on upcoming revisions to the European legislation that is relevant to the provisions of the GDPR and complements the latter. We focus on the three legislative proposals: the AI Act proposal, the Data Governance Act proposal and the forthcoming Data Act proposal.

4.1 AI Act proposal

➢ General comments

The European Commission unveiled a proposal for a new Artificial Intelligence Act (AI Act) in April 2021 (see also Deliverable D2.1 "*Overview & Analysis of the AI Policy Initiatives in EU level*" for a detailed analysis). Not long after the EC proposed the AI Act, the EDPB and the EDPS issued a Joint Opinion 5/2021 on the AI Act proposal. The opinion points out to several critical concerns regarding the wording of the proposal. First, the absence of any reference in the text to the individual affected by the AI system (be they end-users, data subjects or other persons affected by the AI system) appears as a blind spot in the Proposal. The proposal does not address the rights or remedies available to individuals subject to AI systems. Second, the EDPB and EDPS recommend a ban, for both public authorities and private entities, on AI systems categorizing individuals from biometrics (for instance, from face recognition) into clusters according to ethnicity, gender, as well as political or sexual orientation, or other grounds for discrimination (EDPB-EDPS 2021). Accordingly, "biometric categorization" should be prohibited, the opinion concludes. The opinion also calls for a general ban on any use of AI for an automated recognition of human features in publicly accessible spaces - such as of faces but also of gait, fingerprints, DNA, voice, keystrokes and other biometric or behavioral signals - in any context. Moreover, the EDPB and the EDPS consider that the use of AI to infer emotions of a natural person is highly undesirable and should be prohibited, except for certain well- specified use-cases, namely for health or research purposes (e.g., patients where emotion recognition is important).

➢ Interplay with the GDPR

The opinion underlines that given that the development and use of AI systems will in many cases involve the processing of personal data, "ensuring clarity of the relationship of this Proposal to the existing EU legislation on data protection is of utmost importance."

First and foremost, while the recitals of the Proposal clarify that the use of AI systems should still comply with data protection law, the EDPB and EDPS strongly recommend clarifying in Article 1 of the Proposal that the data protection law, in particular the GDPR, shall apply to any processing of personal data falling within the scope of the Proposal (EDPB-EDPS 2021). As explained above (Section 3.1.1.3), in every case the AI system uses personal data, according to the data protection rules, data subjects should always be informed when their data is used for AI training and / or prediction, of the legal basis for such processing, general explanation of the

logic (procedure) and scope of the AI-system. Individuals' right of restriction of processing as well as deletion / erasure of data should always be guaranteed in those cases. Furthermore, the controller should have explicit obligation to inform the data subject of the applicable periods for objection, restriction, deletion of data etc. The AI system must be able to meet all data protection requirements through adequate technical and organizational measures. A right to explanation should provide for additional transparency.

Second, as provided in Section 2.2, different organisations can play different roles under the GDPR in an AI context. The joint Opinion strongly suggests that *"the responsibilities of the various parties - user, provider, importer or distributor of an AI system - need to be clearly circumscribed and assigned".* The clarification is needed with regard to the consistency of these roles and responsibilities with the notions of data controller and data processor carried by the data protection framework since both norms are not correspondent.

As an example, the AI Proposal requires the '*providers'* of the AI system to perform a risk assessment, however, in most cases, the (data) controllers will be the '*users'* rather than providers of the AI systems. For instance, a '*user'* of a facial recognition system is a usually the one deciding on the purposes and means of processing personal data (a '*controller'*) and therefore, is not bound by requirements on AI providers under the AI Proposal). This can create both a confusion and possible gaps regarding the roles and responsibilities of each party involved in the development and deployment of the AI systems. In other words, it is of utmost importance to bring in line the terminology used by the GDPR (data processor, data controller) with the one used by the AI Act (user, provider, distributor).

Third, the EDPB and the EDPS underline that some provisions of the Proposal defining the tasks and powers of the different competent authorities under the AI regulation, their relationships, seem unclear at this stage (EDPB-EDPS 2021). In particular, data protection authorities (DPAs) are already enforcing the GDPR. When AI systems are based on the processing of personal data or process personal data, the DPAs will in such cases assess its compliance with data protection rules. As a result, there will be interconnections of competencies between supervisory authorities under the Proposal and DPAs.

Moreover, scholars have pointed out that how interpretability is distinguished (or not) from the explainability requirement of automated decision-making established in the GDPR is another relevant question. The correlation between the interpretability-transparency requirement in the AI Act and explainability in the GDPR has been considered the example of a non-alignment (Kiseleva 2021).

Lastly, even though the AI Regulation is intended to complement the GDPR, it provides very little clarity on processing of personal data by any other AI system than high-risk AI systems (Bergholm 2021). Data processing of AI systems, which are not high-risk, or related to biometrics or bias, seem to be left solely to the already existing provisions of the GDPR.

➢ Next steps

It is unclear whether and how some of these concerns will be mitigated. The proposal is now being discussed by the co-legislators, the European Parliament and the Council (EU Member states). In November 2021, the Slovenian presidency presented a progress report (draft compromise) on discussions held so far within the Council on the AI draft proposal.[2] Some of the key amendments to the proposal proposed by the Council, include:

- Narrowing the definition of 'AI systems';
- Extending the prohibition of social scoring also to private actors and not merely to public authorities;
- Deleting the necessity to prove the "intention" of behavioural manipulation;
- Extending a list of prohibited practices to the exploitation of vulnerabilities based on social and economic condition of the individual; and, importantly;
- Introducing a broad exception for *"AI systems and their outputs used for the sole purpose of research and development*",

and are therefore bringing the text more in line with the EDPB-EDPS recommendations.

4.2 Data Governance Act proposal

In November 2020, the EC adopted the Proposal for a Data Governance Act (DGA). The DGA proposal consists in the following 3 main pillars. First, the DGA proposal sets the conditions for enhancing the development of the common European data spaces by bringing trust in a range of data sharing services. Second, it introduces a voluntary registration regime applying to 'data altruism' services. Third, the DGA proposal creates a legal regime for the re-use of public sector data which are subject to the rights of third parties.

Both academics (Baloup et al. 2021) and the EDPB-EDPS Joint Opinion 03/2021 highlighted that the DGA entails several significant inconsistencies with the GDPR, notwithstanding the statement in the recital that it is "without prejudice" to the GDPR. The EDPB and the EDPS considered that the Proposal raises significant inconsistencies with the GDPR, as well as with other Union law, in particular as regards the following five aspects: (a) Subject matter and scope of the Proposal (b) Definitions/terminology used in the Proposal; (c) Legal basis for the processing of personal data; (d) Blurring of the distinction between (processing of) personal and non-personal data (and unclear relationship of the Proposal with the Regulation on free flows of non-personal data); (e) Governance/tasks and powers of competent bodies and authorities.

It is beyond the scope of the current deliverable to fully assess these inconsistencies. It suffices to say that in its Statement 05/2021 on the Data Governance Act in light of the legislative developments adopted on 19 May 2021, the EDPB once again urged the co-legislators to carefully consider:

- clarifying the 'interplay' between the DGA and the GDPR ;

[2] https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf

- bringing in line with the GDPR the definitions/terminology used in the DGA ;
- specifying whether the provisions of the DGA refer to non-personal data, personal data or both, and also specify that in case of 'mixed data sets' the GDPR applies.

On November 30, 2021, the EU Parliament and Council reached a provisional agreement on the proposed Data Governance Act. There are the following points worth mentioning. First, as already said, the DGA creates a framework to foster a new business model – data intermediation services**.** These services will support voluntary data-sharing between companies or facilitate the fulfilment of data-sharing obligations set by law. The proposal foresees that it will also help people to have control over their data and allow them to share it with a company they trust. This can be done, for example, by means of novel personal information management tools, such as personal data spaces or data wallets, which are apps that share such data with others, based on the data holder's consent. The service providers will not be allowed to use shared data for other purposes. They will not be able to benefit from the data – for example, by selling it on. Second, the DGA also makes it easier for individuals and companies to make data voluntarily available for the common good, such as medical research projects.

Entities seeking to collect data for objectives of general interest may request to be listed in a national register of recognised data altruism organisations. Registered organisations will be recognised across the EU. The legislators hope to create the necessary trust in data altruism, encouraging individuals and companies to donate data to such organisations so that it can be used for the wider societal good.

4.3 Data Act proposal

The Data Act is intertwined with the Data Governance Act, within the scheme of the European Data Strategy. Although the proposed Act's text has not been published by the EC yet, in May 2021 the EC published the Inception Impact Assessment on the upcoming Data Act. The initiative would look both at data usage rights in industrial value chains and particularly at a fair distribution of usage rights that allow all parties to benefit from data-driven innovation. The Data Act would aim to enhance clarity on the rules with respect to B2B access to and sharing of data, both non-personal and personal, by ensuring in particular data can be shared safely and not misappropriated, and in line with the applicable EU legislation, including the GDPR. The initiative in this respect seeks to provide a coordinated response that takes into account existing legal instruments such as the General Data Protection Regulation, the ePrivacy Directive and the Trade Secrets Directive, as well as the Database Directive, which could be amended so that it supports the objectives of this initiative.

The proposal was planned to be out in the last quarter of 2021. However, the Regulatory Scrutiny Board, an independent body that quality-checks the Commission's impact assessment for new legislative proposals, rejected the Data Act proposal in October 2021. At the moment of writing, it is expected that the draft Data Act proposal will be presented on 23 of February 2022.

5. Recommendations

5.1 Conclusion: existing gaps and challenges

➢ The lack of common definitions and formalism about reliable AI

Defining a formal framework for reliability, transparency and fairness in AI is currently a strong need. The inconsistencies in the terminology reported in Section 3.1.1 show that terms such as interpretable, explainable and transparent convey different meanings and are 'weighted' differently in the technical and social sciences. The adoption and development of reliable AI practices is hindered by the lack of an overarching framework that is understood and used by regulators, sociologists, psychologist, ethicists and technicians. As a step towards a common formalism to define trusted AI, we propose in the following a multidisciplinary definition of interpretable AI that may be adopted in both the social and the computer sciences.

In daily language, an object is defined as interpretable if it is possible to find its interpretation, hence if we can find its meaning. A formal definition of interpretability exists in the field of mathematical logic, and it can be summarized as the possibility of interpreting, or translating, one formal theory into another while preserving the validity of each theorem in the original theory during the translation. The translated theory as such assigns meaning to the original theory and it is an interpretation of it. The translation may be needed, for instance, to move into a simplified space where the original theory is easier to understand and can be presented in a different language.

From these explicit definitions we can derive the following, multidisciplinary definition of interpretability that embraces both technical and social aspects:

> "Interpretability is the capability of assigning meaning to an instance by a translation that does not change its original validity".

The definition of interpretable AI can be then derived by clarifying what should be translated:

> "An AI system is interpretable if it is possible to translate its working principles and mechanisms in human-understandable language without affecting the validity of the system".

Interpretability is needed to make the output generation process of an AI system explainable and understandable to humans.

The definition that we provide clarifies that interpretability is often obtained as a translation process. Such process may be introduced directly at the design stage as an additional task of the system. If not available by design, interpretability may be obtained by post-hoc explanations that aim at improving the understandability of how the outcome was generated. Interpretability can thus be sought through iterations and in multiple forms (e.g., graphical visualizations,

natural language, or tabular data), which can be adapted to the receiver. This fosters the auditability and accountability of the system.

- Diverging legal terminology

The second challenge identified is the diverging terminology and definitions of explainability, explicability, and transparency in various policy and legislative documents. As explained above in Section 3.1.1.3, the GDPR distinguished between legally binding transparency requirements concerning personal data processing and non-binding Recital 71, which suggests the 'explainability' of automated-decision making.

The draft AI Act envisions explainability as part of transparency, the latter depending heavily on the types of AI systems defined in the draft regulation. This is reinforced by Recital 47, which directly connects the complexity and opacity of certain AI systems with the need for transparency. This is also a justification for the requirement in Article 13 for high-risk AI systems to be *"designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately."*

The obvious difference here, in comparison with the idea in the AI HLEG Guidelines, is that the transparency and hence the explainability are addressed towards the 'users', which is a category that has been legally defined in Article 3 (4) of the AI Act meaning ''any natural or legal person, public authority, agency or other body using AI systems under its authority, except when the AI system is used in the course of a personal non-professional activity." This means that people (e.g., end-users) who are in some way adversely impacted by an AI system would not necessarily have the means to find out or prove it since the explainability and the transparency obligations are not addressed towards them.

An exception of this are three types of AI systems defined in Article 52 of the draft AI Act. For AI systems that interact with natural persons, emotion recognition systems or a biometric categorisation systems and AI systems that generate deep fakes, the AI Act proposal prescribes obligation for an additional layer of transparency in the form of disclosure that would make people aware they interact or are exposed to such systems. It is interesting that even though the AI Act proposal does use the very term transparency, in these particular cases it does not encompass the explainability and the traceability dimension that were part of the concept according to the AI HLEG Guidelines.

This shows the inconsistence of the terminology from a legal point of view, which could be partially mended by the amendments to the draft AI Act but if not, would be subject to interpretation by the Court of Justice of the European Union, the later relying on the interpretation of other branches of science to complement the legal gaps, which shows the clear necessity of unified taxonomy.

- The incomputability

Another challenge lies in what Hildebrandt coined as *'the foundational incomputability of human identity'*; meaning that any computation of our interactions can be performed in multiple ways

— leading to a plurality of potential identities (Hildebrandt 2019). Building on this, privacy is not computable and one cannot formalize privacy completely. There are very different concepts of privacy and the concepts depend on the environment, the circumstances, the jurisdiction etc.

Legal norms are inherently text-driven and language-based and anchored in the semantic ambiguities of natural language. The high contextualization and ambiguity of natural language – and therefore legal norms – is, however, not a bug but it is a feature.

A code-driven world demands to formalise whatever requirements to be translated into code. Formalisation enables the logical operation of deduction, in the sense of 'if this then that' (IFTTT). Such operations are crucial for automation, which is the core of computing systems. To the extent that formalisation is not possible or questionable, code-driven architectures cannot be developed or may be unreliable. The second constraint is the need to disambiguate the terms used when formulating the requirements. This constraint is in turn inherent in formalisation, because deduction is not possible, if it remains unclear what the precise scope of the requirements is. Disambiguation implies an act of interpretation that should result in a clear demarcation of the consequences of applying the relevant terms (Hildebrandt 2019).

The changing circumstances that may destabilise common sense interpretations of legal norms. The terms of a contract may seem clear and distinct, but in the case of unexpected events a reasonable interpretation may unsettle mutual expectations and require their reconfiguration. Text-driven law is adaptive in a way that would be difficult to achieve in code driven law (which relies on a kind of completeness that is neither attainable nor desirable).

These finding apply to notions such as privacy, explainability or fairness. Hildebrandt sums up that 'translating legal or ethical notions of fairness into machine learning research design is not at all obvious, also because different notions of fairness may be incompatible.'

5.2 Ways forward

In this section, we present a preliminary selection of ways forward based on the observations done for this research. A common remark to these ways forward is the need to ensure dialogue between the different stakeholders involved in AI systems from the design, the use, and the enforcement perspective in order to provide the most complete guidance to fill in the gaps identified. Indeed, industry, researchers, data subjects' associations, regulatory and enforcement authorities should all dialogue for ensuring the higher level of trusted AI (Kuziemski and Palka 2019). Algorithmic explainability and transparency must be catered to their audience and supporting individual's understanding of algorithmic processes and enabling informed decisions. This can only be achieved by involving users and experts in the design and development process of the AI systems.

As outlined by some authors, one important aspect of any ways forward is to solve the power asymmetries for AI development with limited private companies' monopoly over data and to develop AI inclusive policies and regulations (Kuziemski and Palka 2019). This can be achieved by putting citizen-centred innovation with class action, activism, and whistle-blowers schemes.

Opening up decision-making processes through APIs, well-designed research exceptions could also be part of the improvement solutions.

In addition, providing information on the use of clean data sets is key to achieve the purpose of trusted AI. This includes information such as “identifiable and retraceable origins, provably legally and legitimately obtained, in compliance to the GDPR and other relevant regulations, with due regard for any risks of excessive bias, discrimination and prejudice” (Lambrecht et al., n.d.)

All the ways forward detailed below are collectively working towards more trusted AI. Figure 10 presents an overview of the ways forward.

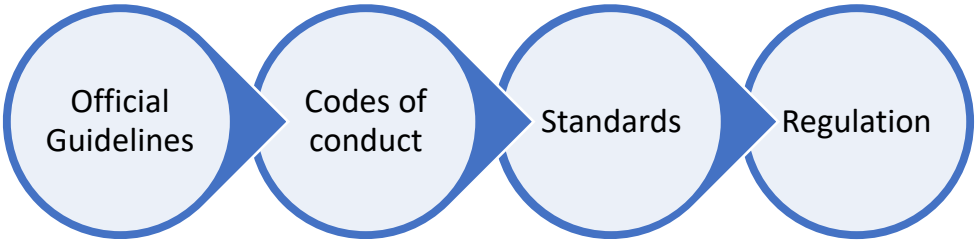

Figure 10: Overview of the ways forward.

- Official guidelines on AI and GDPR

Firstly, as observed by many researchers, scholars, practitioners and data subjects, there are uncertainties on how to interpret some GDPR provisions in an AI context. This requires close collaboration with developers, designers, computer scientists, ethics and legal practitioners. AI system uptake is growing and in order to provide legal certainty and ensure a trustworthy use of AI, further guidance and clarification on the interpretation of the GDPR provisions is more than welcome. Some clarifications are coming through in various guidelines provided by the EDPB or the EDPS but a coherent guidance from an authoritative source is needed for ensuring collective understanding of the GDPR. This could possibly materialise in the future through the Court of Justice of the European Union jurisprudence.

- Codes of conduct

Further guidance can be materialised under the format of Codes of Conduct, which are especially useful to translate the legal principles and obligations in concrete recommendations for designers and developers on how to enforce and ensure compliance from the earliest stage of an AI system design.

Some examples of codes of conduct could be:

- on the use of clean data sets;
- on the information provided to users, for example, about how to communicate the multiple information at different levels of abstraction;
- on best practices to deal with data subjects' right requests;
- on mapping and making available best practices and technological solutions to ensure AI conformity with GDPR.

➢ Standards

To ensure trustworthy development of AI systems in line with the GDPR provisions, specific standards could be part of the solutions too. Especially, to provide practical solutions and formats according to which AI systems could be designed in a GDPR-compliant way. This could include interfaces for data subjects that could have various compliance purposes such as transparency, interaction, data subjects' right enforcement and so forth. For instance, an opting in and opting out option should be accessible at the same level of difficulty. Seamingful design can transform perceived flaws into increased understanding and mitigate dissatisfaction when a system does not work as expected (Lambrecht et al., n.d.). Seamingful is a concept defined by Eslami and others and can be defined as a design that makes "system infrastructure elements visible when the user actively chooses to understand or modify that system" (Eslami et al. 2016).

The industry has also started to work on standardized formats of providing an information to a user. IBM's AI Factsheets 360 (https://aifs360.mybluemix.net/introduction) are an example of an initiative that aims at providing better transparency for informed usage. Information is provided to help users understand how a model was generated, and thus, to determine whether its usage is appropriate for a given task.

➢ Regulation

The clarification could also come from the legislative level, strong of the research conducted on the GDPR and the development of AI. Perhaps a mix between global and sectoral regulation revision or adoption could bring the clarification needed. Thus, creation of a checks and balances system where algorithms and/or datasets may be subject to scrutiny by an independent (self-/co-) regulatory body on their compliance to the above codes and regulations. However, this seems hard to enforce in practice, given the considerable amount of algorithms and datasets.

6. Conclusions

In this initial version of this deliverable, we observe how the GDPR framework applies in an AI system context and which aspects remain unclear and undermine the uptake of AI.

The following general takeaways from Section 3 are worth underlining:

1) **The GDPR follows a technology-agnostic approach.** It does not at any points refer to 'artificial intelligence', nor any terms expressing related concepts, such as intelligent systems, autonomous systems, automated reasoning and inference, machine learning or even big data. This does not, however, mean that the GDPR does not apply to training, testing, validation or deploying the AI systems.
2) **The GDPR applies to personal data.** Although it can be argued that the training data often consist of non-personal information, in case of a doubt it is best to assume that data are personal data. This is because true anonymization is a very onerous standard. Due to a risk for re-identification, having gone through an anonymization process at a certain point in time should not be viewed as a silver bullet for circumventing the application of the GDPR.
3) **The GDPR addresses data controllers.** The GDPR addresses controllers, not computer scientists, but depending on the circumstances, computer scientists, system developers or any other individual or organization can be considered a "data controller". In Table 1, we presented different roles an organization can play depending on the lifecycle of AI system development.
4) **The extent to which the GDPR applies to AI is being (re)defined.** The guidance from official institutions such as the EDPB and the EDPS focuses only in part on the AI systems. There is a lack of sufficient clarity, uncertainties and diverging opinions between scholars and interpretative guidelines. The academic literature showed sometimes converging and in other cases conflicting opinions among the research community on the scope of some GDPR provisions applied to AI systems.
5) **Compliance with data subjects' rights is a growing challenge.** GDPR's provisions impose obligations on controllers and processors to make sure they comply with data subjects' rights. As demonstrated, there are a lot of uncertainties of how to ensure the compliance with the right to erasure, right of access, right to data portability etc. in the context of AI. Undoubtedly, there is a growing movement both from the individuals themselves, and from civil society organizations to enforce the GDPR rights. The best illustration is the recent consent pop up decision against IAB and its illegal advertising practices.

Moreover, the analysis in Section 4 hints on the fact that the European legislator is well aware of some of these legal challenges. The new initiatives, such as the AI Act proposal, the Data Governance Act proposal and the forthcoming Data Act proposal promise to create additional safeguards, data quality requirements and favourable conditions to enhance data sharing. These upcoming legislative initiatives will not replace or considerably affect the GDPR, but will

rather have a complementary role. At this stage it is however unclear how they will develop throughout a legislative process and how they will be interpreted.

Section 5 provided initial conclusions on the existing gaps and challenges observed. It also suggested ways forward to address them.

D4.3 is the initial version of the analysis of the legal and ethical framework of trusted AI. The version will therefore be updated and improved for its final version (D4.4 – Final analysis of the legal and ethical framework of trusted AI, to be delivered in M36). Based to this first part of the research, we will continue the comprehensive analysis of the legal data protection framework for the use of AI applications. The future research direction includes applying the applicable provisions to AI systems used in media environments. The gaps identified will be further researched on, focusing on how they can be solved through a revision of the legislation. We will also research best practices about how people can be aware of what is done with their data in the media environment.

7. References

Article 29 Working Party. 2010. ' Opinion 3/2010 on the principle of accountability' https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp173_en.pdf

Article 29 Data Protection Working Party. 2013. 'Opinion 03/2013 on Purpose Limitation'.

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'. WP217. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

Article 29 Data Protection Working Party. 2014. 'Opinion 05/2014 on Anonymisation Techniques'.

Article 29 Data Protection Working Party. 2018. 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679'.

Article 29 Working Party. 2018. Guidelines on consent under Regulation 2016/679

Article 29 Working Party. 2018. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679

Baloup, Julie, Emre Bayamlıoğlu, Aliki Benmayor, Charlotte Ducuing, Lidia Dutkiewicz, Teodora Lalova, Yuliya Miadzvetskaya, and Bert Peeters. 2021. 'White Paper on the Data Governance Act'. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3872703.

Bergholm, Jenny. 2021. 'The GDPR and the Artificial Intelligence Regulation – It Takes Two to Tango? - CITIP Blog'. n.d. https://www.law.kuleuven.be/citip/blog/the-gdpr-and-the-artificial-intelligence-regulation-it-takes-two-to-tango/.

Biasin, Elisabetta. 2021. 'Why Accuracy Needs Further Exploration in Data Protection'. In *Proceedings of the 1st International Conference on AI for People: Towards Sustainable AI*, 1–7. EAI. https://doi.org/10.4108/eai.20-11-2021.2314205.

Bibal, Adrien, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. 2021. 'Legal Requirements on Explainability in Machine Learning'. Artificial Intelligence and Law 29 (2): 149–69. https://doi.org/10.1007/s10506-020-09270-4.

Biega, Asia J., and Michèle Finck. 2021. 'Reviving Purpose Limitation and Data Minimisation in Data-Driven Systems'. *Technology and Regulation*, August, 44-61 Pages. https://doi.org/10.26116/techreg.2021.004.

Binns, Reuben. 2019. 'Enabling Access, Erasure, and Rectification Rights in AI Systems'. *Information Commissioner's Office*, 15 October 2019. https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/.

Chen, Jiahong. 2018. 'The Dangers of Accuracy': *European Data Protection Law Review* 4 (1): 36–52. https://doi.org/10.21552/edpl/2018/1/7.

Clifford, Damian, and Jef Ausloos. 2018. 'Data Protection and the Role of Fairness'. Yearbook of European Law 37 (January): 130–87. https://doi.org/10.1093/yel/yey004.

Dignum, Virginia. 2021. 'The Myth of Complete AI-Fairness'. *ArXiv:2104.12544 [Cs]*, April. http://arxiv.org/abs/2104.12544

Dimitrova, Diana. 2021. 'The Rise of the Personal Data Quality Principle. Is It Legal and Does It Have an Impact on the Right to Rectification?' SSRN Scholarly Paper ID 3790602.

Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3790602.
Edwards, Lilian, and Michael Veale. 2017. 'Slave to the Algorithm? Why a "right to an Explanation" Is Probably Not the Remedy You Are Looking For'. Preprint. LawArXiv. https://doi.org/10.31228/osf.io/97upg.
Emanuilov, Ivo, Stefano Fantin, Plixavra Vogiatzoglou, and Thomas Marquenie. n.d. 'Purpose Limitation By Design As A Counter To Function Creep And System Insecurity In Police Artificial Intelligence'. *UNICRI Special Collection on AI in Criminal Justice*, August 2020.
Eslami, Motahhare, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. 'First I "like" It, Then I Hide It: Folk Theories of Social Feeds'. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2371–82. CHI '16. San Jose, California, USA: Association for Computing Machinery. https://doi.org/10.1145/2858036.2858494.
European Commission. 2019. ' Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union' . https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52019DC0250&rid=2
European Data Protection Board. 2018. Guidelines 05/2020 on consent under Regulation 2016/679
European Data Protection Board. 2021. Guidelines 02/2021 on virtual voice assistants
European Data Protection Board. 2022. 'Guidelines 01/2022 on Data Subject Rights - Right of Access Version 1.0'. https://edpb.europa.eu/our-work-tools/documents/public-consultations/2022/guidelines-012022-data-subject-rights-right_en.
EDPB-EDPS. 2021. EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) https://edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf
European Data Protection Supervisor. 2020. 'A Preliminary Opinion on Data Protection and Scientific Research'. https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf.
Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI'. SSRN Scholarly Paper ID 3518482. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3518482.
Fouad, Imane, Cristiana Santos, Feras Al Kassar, Nataliia Bielova, and Stefano Calzavara. 2020. 'On Compliance of Cookie Purposes with the Purpose Specification Principle'. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*, 326–33. https://doi.org/10.1109/EuroSPW51379.2020.00051.
Goldsteen, Abigail, Gilad Ezov, Ron Shmelkin, Micha Moffie, and Ariel Farkash. 2021. 'Data Minimization for GDPR Compliance in Machine Learning Models'. *AI and Ethics*, September. https://doi.org/10.1007/s43681-021-00095-8.
Goodman, Bryce, and Seth Flaxman. 2016. 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"'. ArXiv:1606.08813 [Cs, Stat], June. http://arxiv.org/abs/1606.08813.
Graef, Inge. 2020. 'THE OPPORTUNITIES AND LIMITS OF DATA PORTABILITY FOR STIMULATING COMPETITION AND INNOVATION'. Competition Policy International, 10.

Hacker, Philipp. 2021. 'A Legal Framework for AI Training Data—from First Principles to the Artificial Intelligence Act'. Law, Innovation and Technology, September, 1–45. https://doi.org/10.1080/17579961.2021.1977219.

Hallinan, Dara, and Frederik Borgesius. 2020. 'Opinions Can Be Incorrect (in Our Opinion)! On Data Protection Law's Accuracy Principle'. *International Data Privacy Law* 10 (February): 1–10. https://doi.org/10.1093/idpl/ipz025.

Hamon, Ronan, Henrik Junklewitz, Gianclaudio Malgieri, Paul De Hert, Laurent Beslay, and Ignacio Sanchez. 2021. 'Impossible Explanations?: Beyond Explainable AI in the GDPR from a COVID-19 Use Case Scenario'. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 549–59. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445917.

Hestness, Joel, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. 'Deep Learning Scaling Is Predictable, Empirically'. *ArXiv:1712.00409 [Cs, Stat]*, December. http://arxiv.org/abs/1712.00409.

High-Level Expert Group on Artificial Intelligence. 2019. 'Ethics Guidelines for Trustworthy AI'.

Hildebrandt, Mireille. 2019. 'Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning'. Theoretical Inquiries in Law 20 (1): 83–121. https://doi.org/10.1515/til-2019-0004.

Hill, Kashmir, Krolik, Aaron 'How Photos of Your Kids Are Powering Surveillance Technology - The New York Times'. n.d. https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html.

Hoboken, Joris van. 2016. 'From Collection to Use in Privacy Regulation? A Forward Looking Comparison of European and U.S. Frameworks for Personal Data Processing'. In *Exploring the Boundaries of Big Data*, 231–59. Netherlands Scientific Council for Government Policy.

Hoeren, Thomas. 2017. 'Big Data and the Legal Framework for Data Quality'. *International Journal of Law and Information Technology* 25 (1): 26–37. https://doi.org/10.1093/ijlit/eaw014.

Hu, Yipeng, Joseph Jacob, Geoffrey J. M. Parker, David J. Hawkes, John R. Hurst, and Danail Stoyanov. 2020. 'The Challenges of Deploying Artificial Intelligence Models in a Rapidly Evolving Pandemic'. *Nature Machine Intelligence* 2 (6): 298–300. https://doi.org/10.1038/s42256-020-0185-2.

ICO 'Guidance on AI and Data Protection'. 2021. https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/.

Jasserand, Catherine. 2018. 'Massive Facial Databases and the GDPR: The New Data Protection Rules Applicable to Research'. In Data Protection and Privacy: The Internet of Bodies, 169–88. Hart Publishing/Bloomsbury Publishing Plc.

Karst, Kenneth L. 1966. 'The Files': Legal Controls over the Accuracy and Accessibility of Stored Personal Data'. *Law and Contemporary Problems* 31 (2) (4).

Kindt, Els. 2007. 'Biometric Applications and the Data Protection Legislation'. *Datenschutz Und Datensicherheit - DuD* 31 (March): 166–70. https://doi.org/10.1007/s11623-007-0064-6.

Kiseleva, Anastasiya. 2021. 'COMMENTS ON THE EU PROPOSAL FOR THE ARTIFICIAL INTELLIGENCE ACT'. SSRN Scholarly Paper ID 3949585. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3949585.

Knowledge Centre Data & Society, "Artificial Intelligence and Data Protection: An Exploratory Guide", October 2021'. n.d.

Koops, Bert-Jaap. 2011. 'The (In)Flexibility of Techno-Regulation and the Case of Purpose-Binding'. *Legisprudence* 5 (2): 171–94.

Koops, Bert-Jaap. 2020. 'The Concept of Function Creep'. SSRN Scholarly Paper ID 3547903. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3547903.

Kuebler-Wachendorff, Sophie, Robert Luzsa, Johann Kranz, Stefan Mager, Emmanuel Syrmoudis, Susanne Mayr, and Jens Grossklags. 2021. 'The Right to Data Portability: Conception, Status Quo, and Future Directions'. Informatik Spektrum 44 (4): 264–72. https://doi.org/10.1007/s00287-021-01372-w.

Kuziemski, Maciej, and Przemyslaw Palka. 2019. *AI Governance Post-GDPR : Lessons Learned and the Road Ahead*. European University Institute. https://doi.org/10.2870/470055.

Lambrecht, Ingrid, Elias Storms, Aleksandra Kuczerawy, and David Geerts. n.d. 'Algorithmic Transparency and Accountability in Practice (ATAP) - WP2-T.2 Policy and UI Guidelines'. KU Leuven Centre for IT and IP law and the Meaningful interactions Lab (Mint Lab).

MacCarthy, Mark. 2018. 'In Defense of Big Data Analytics'. SSRN Scholarly Paper ID 3154779. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3154779.

Malgieri, Gianclaudio. 2020. 'The Concept of Fairness in the GDPR: A Linguistic and Contextual Interpretation'. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 154–66. Barcelona Spain: ACM. https://doi.org/10.1145/3351095.3372868.

Mayer-Schönberger, Viktor. 2016. 'Regime Change? Enabling Big Data Through Europe's New Data Protection Regulation'. *Columbia Science & Technology Law Review* 17 (April): 315.

Moerel, Lokke, and Corien Prins. 2016. 'Privacy for the Homo Digitalis: Proposal for a New Regulatory Framework for Data Protection in the Light of Big Data and the Internet of Things'. SSRN Scholarly Paper ID 2784123. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.2784123.

Murgia, Madhumita. 2019. 'Microsoft Quietly Deletes Largest Public Face Recognition Data Set | Financial Times'. https://www.ft.com/content/7d3e0d6a-87a0-11e9-a028-86cea8523dc2.

Naudts, Laurens, Ingrid Lambrecht, Pierre Dewitte, Jef Ausloos, Oscar Alvarado, and Jeroen Wauman. n.d. 'Algorithmic Transparency and Accountability in Practice (ATAP) - Mapping Legal and HCI Scholarship Interdisciplinary Problem Formulation'. KU Leuven Centre for IT and IP law and the Meaningful interactions Lab (Mint Lab).

Noyb. 2018. 'GDPR: Noyb.Eu Filed Four Complaints over "Forced Consent" against Google, Instagram, WhatsApp and Facebook'', 25 May 2018. https://noyb.eu/wp-content/uploads/2018/05/pa_forcedconsent_en.pdf.

Opinion 1/15 of the Court (Grand Chamber). 2017, ECLI:EU:C:2017:592. Court of Justice of the European Union.

Papernot, Nicolas, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. 2017. 'Semi-Supervised Knowledge Transfer for Deep Learning from Private Training Data'. ArXiv:1610.05755 [Cs, Stat], March. http://arxiv.org/abs/1610.05755.

Pierson, J., Robinson, S. C., Boddington, P., Chazerand, P., Kerr, A., Milan, S., ... Aconstantinesei, I. C. (2021). AI4People - AI in Media and Technology Sector: Opportunities, Risks, Requirements and Recommendations. Brussels: Atomium - European Institute for Science, Media and Democracy (EISMD).

Raji, Inioluwa Deborah, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. 'Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing'. ArXiv:2001.00964 [Cs], January. http://arxiv.org/abs/2001.00964.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. '"Why Should I Trust You?": Explaining the Predictions of Any Classifier'. ArXiv:1602.04938 [Cs, Stat], August. http://arxiv.org/abs/1602.04938.

Rocher, Luc, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. 2019. 'Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models'. *Nature Communications* 10 (1): 3069. https://doi.org/10.1038/s41467-019-10933-3.

Sartor, Giovanni, European Parliament, European Parliamentary Research Service, and Scientific Foresight Unit. 2020. *The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence: Study*. http://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf.

Scott, Mark, Manancourt, Vincent 2021 'Google and Data Brokers Accused of Illegally Collecting People's Data: Report' 2020. https://www.politico.eu/article/google-and-data-brokers-accused-of-illegally-collecting-data-report/

Selbst, Andrew D, and Julia Powles. 2017. 'Meaningful Information and the Right to Explanation'. International Data Privacy Law 7 (4): 233–42. https://doi.org/10.1093/idpl/ipx022.

Senarath, Awanthika, and Nalin Asanka Gamagedara Arachchilage. 2018. 'Understanding Software Developers' Approach towards Implementing Data Minimization'. *ArXiv:1808.01479 [Cs]*, August. http://arxiv.org/abs/1808.01479.

Shanmugam, Divya, Samira Shabanian, Fernando Diaz, Michèle Finck, and Asia Biega. 2021. 'Learning to Limit Data Collection via Scaling Laws: Data Minimization Compliance in Practice'. *ArXiv:2107.08096 [Cs]*, July. http://arxiv.org/abs/2107.08096.

Sun, Chen, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. 'Revisiting Unreasonable Effectiveness of Data in Deep Learning Era'. In , 843–52. https://openaccess.thecvf.com/content_iccv_2017/html/Sun_Revisiting_Unreasonable_Effectiveness_ICCV_2017_paper.html.

Ustun, Berk, and Cynthia Rudin. 2016. 'Supersparse Linear Integer Models for Optimized Medical Scoring Systems'. Machine Learning 102 (3): 349–91. https://doi.org/10.1007/s10994-015-5528-6.

Veale, Michael, Reuben Binns, and Lilian Edwards. 2018. 'Algorithms That Remember: Model Inversion Attacks and Data Protection Law'. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2133): 20180083. https://doi.org/10.1098/rsta.2018.0083.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2020. 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI'. *SSRN Electronic Journal*.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation'. International Data Privacy Law 7 (2): 76–99. https://doi.org/10.1093/idpl/ipx005.

Wen, Hongyi, Longqi Yang, Michael Sobolev, and Deborah Estrin. 2018. 'Exploring Recommendations under User-Controlled Data Filtering'. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 72–76. RecSys '18. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3240323.3240399.

Wisman, T.H.A. 2013. 'Purpose and Fuction Creep by Design: Transforming the Face of Surveillance through the Internet of Things'. *European Journal of Law and Technology* 2013 (2).

Wong, Janis, and Tristan Henderson. 2019. 'The Right to Data Portability in Practice: Exploring the Implications of the Technologically Neutral GDPR'. International Data Privacy Law 9 (3): 173–91. https://doi.org/10.1093/idpl/ipz008.

AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu