Ethics in Artificial Intelligence (Module 2)

Last update: 13 April 2025

Academic Year 2024 – 2025
Alma Mater Studiorum · University of Bologna

Contents

1 AI in the GDPR

Remark (AI risks)**.**

- Eliminate or devalue jobs.
- Lead to poverty and social exclusion, if no measures are taken.
- Concentrate economic wealth in a few big companies.
- Allow for illegal activities.
- Surveillance, pervasive data collection, and manipulation.

  Example. Many platforms operate in a two-sided market where users are on one side and advertisers, the real source of income, are on the other.
- Public polarization and interference with democratic processes.
- Unfairness, discrimination, and inequality.
- Loss of creativity.

  Remark. Creativity can be:

  Combinatorial Combination of existing creativity.

  Exploratorial Explore new solutions in a given search space.

Remark. In the GDPR, there are no references to artificial intelligence.

1.1 Introduction

1.1.1 Definitions (article 4)

Personal data

Personal data Any information relating to an identified or identifiable natural person (the data subject). It excludes information that are not related to humans (e.g., natural phenomena) or that do not refer to a particular individual (e.g., information on human physiology or pathologies).

Natural person Individual person (i.e., not companies, which are legal persons).

Identifiable natural person Person that can be identified directly or indirectly using, for instance, name, username, identifier (e.g., in pseudonymization), physical features, economic status, ...

Remark. The GDPR does not contain a positive definition of non-personal data. Anything that is not considered personal data is non-personal.

Processing

Processing Any operation performed on personal data either manually or using automated systems.

Controller

Controller Natural or legal person, public authority, agency, or other bodies which determines the purposes and means for processing personal data.

Processor Natural or legal person, public authority, agency, or other bodies that processes personal data on behalf of a controller. Processor

1.1.2 Territorial scope (article 3)

The GDPR applies to the processing of personal data whenever:

- The controller or processor resides in the EU, regardless of where processing physically takes place.
- The data subject (of any nationality) is in the EU, regardless of where the controller or processor resides, when the purpose is for:
  - Offering goods or services, independently of whether a payment is required.
  - Monitoring of behavior.

1.2 Data protection principles

1.2.1 Lawfulness of processing (article 6)

Processing of personal data is lawful if at least one of the following conditions applies: Lawfulness of processing

Consent The data subject has given consent to process its personal data for some specific purposes.

Necessity Processing personal data is necessary for a certain aim. This applies when:

- Processing is necessary prior to entering a contract or for the performance of the contract itself the data subject is part of.

  Example. Before concluding the contract for an insurance, the insurer is allowed to process personal data to determine the premium.

  Example. When using a delivery app, processing the address without asking anything is lawful.

- Processing is necessary for compliance with legal obligations the controller is subject to.

  Example. Companies have to keep track of users' purchases in case of tax inspection.

- Processing is necessary to protect the vital interests of the data subject or another natural person.

  Example. The medical record of an unconscious patient can be accessed by the hospital staff.

- Processing is necessary to perform a task carried out in the public interest.

  Example. Processing personal data for public security is allowed.

Legitimate interest Processing is necessary to pursue the controller's legitimate interests, unless overridden by the interests and fundamental rights of the data subject.

Remark. As a rule of thumb, legitimate interests of the controller can be pursued if only a reasonably limited amount of personal data is used.

Example. The gym one is subscribed in can send (contextual) advertisements by email to pursue economic interests.

Remark. Targeted advertising is in principle prohibited. However, companies commonly pair legitimate interest with the request for consent.

1.2.2 Transparency (article 5)

Transparency

Any information regarding data processing (e.g., privacy policy) addressed to the public or to the data subject should be concise, accessible, and easily understandable.

1.2.3 Fairness (article 5)

Informational fairness

Informational fairness Data subjects should be informed of the existence of data processing and profiling, and its purposes. Controllers should provide the data subject with any further information needed to ensure fairness, transparency, and accountability.

Substantive fairness

Substantive fairness Controllers should implement measures to correct inaccuracies, minimize risks, and secure sensitive personal data.

1.2.4 Purpose limitation (article 5)

Purpose limitation

The personal data collected should be for a specified, explicit, and legitimate purpose. Further processing for incompatible purposes is not allowed, unless it is for archiving purposes in the public interest, scientific or historical research, and statistical purposes. Criteria to determine whether repurposing is compatible are:

- The distance between the new and original purpose,
- The alignment of the new purpose with the data subject's expectations, the nature of the data (e.g., if the data is related to protected categories), and their impact on the data subject's interests,
- The measures adopted by the controller to guarantee fairness and prevent risks.

Remark. When the data is used for compatible purposes not foreseen when the data was collected, the data subject should be informed.

Remark. Putting the data subject's anonymized data into the training set of a model is allowed as the trained model as-is does not directly affect them.

1.2.5 Data minimization (article 5)

Data minimization

Data collected from the data subject should be adequate, relevant, and limited with respect to the purpose it is required for.

Remark. Data minimization does not imply that additional data cannot be collected, as long as the benefits outweigh the risks.

Remark. Minimization is less strict for statistical purposes as they do not target specific individuals.

1.2.6 Accuracy (article 5)

Accuracy

Personal data related to an individual should be accurate and kept up to date. Inaccuracies for the purpose the data was collected for must be rectified or erased.

1.2.7 Storage limitation (article 5)

Personal data should be kept only for the time needed for its purpose. Longer storage is allowed for archiving, research, and statistical purposes. Storage limitation

1.3 Personal data (article 4.1)

1.3.1 Identifiability

Identifiability Condition under which some data not explicitly linked to a person allows to still identify that person. Identifiability

In this case, the data that allows re-identification is considered personal data.

> **Remark.** The identifiability of some data depends on the current technological and sociotechnical state-of-the-art (i.e., if it takes a lot of time to re-identify, it does not count as personal data).

Pseudonymization Substitute data items identifying a person with pseudonyms. The link between pseudonym and real data can be traced back. Pseudonymization

Anonymization Substitute data items identifying a person with (in theory) non-linkable information. Anonymization

> **Remark.** Re-identification is usually performed using statistical correlation between anonymized data and other sources.
> With statistical methods, re-identified data is considered personal data as long as there is a sufficient degree of certainty.

> **Example.** There are many cases of anonymized datasets that have been re-identified, for instance:
>
> - Journalists were able to re-identify politicians based on a browsing history dataset.
> - Researchers were able to re-identify anonymized medical records.
> - Anonymized ratings in the Netflix price database were traced back to their authors in IMDb.

1.3.2 Inferred data

Inferred personal data New information about a data subject obtained using algorithmic models on its personal data. Inferred personal data

> **Remark.** There are two cases about inferred data presented to the European Court of Justice:
>
> 1. Related to the application for a residence permit, the Court stated that only the provided data and the final conclusion are personal data, while intermediate conclusions are not.
> 2. In a subsequent case, related to an exam script, the Court stated that the examiner's comments (i.e., data inferred from the data subject's exam) are to be considered personal data.

> **Remark.** According to the European Data Protection Board, inferred data are considered personal data. However, some rights do not apply.

Example. In an exam, the comments of an examiner are inferred data. However, the data subject does not have the right to rectification (unless there is a mistake from the examiner).

Remark. When personal data are embedded into an AI system through training, they are not considered personal data anymore. Only when performing inference the output is again personal data.

Right to "reasonable inference" Right that is currently under discussion.

Right to "reasonable inference"

It is the right to have decisions affecting data subjects performed using reasonable inference systems that respect ethical and epistemic standards.

Remark. Data subjects should have the right to challenge the results of inference, and not only the final decision based on inferred data.

Remark. Inference can be unreasonable if it does not affect data subjects (e.g., for research purposes).

Reasonable inference has the following criteria:

Acceptability Input data for inference should be normatively acceptable for their final purpose (e.g., ethnicity cannot be used to infer whether an individual is a criminal).

Relevance The inferred information should be normatively acceptable for their final purpose (e.g., ethnicity cannot be inferred from the available data if the purpose is for approving a loan).

Reliability Input data, training data, and processing methods should be accurate and statistically reliable.

1.4 Profiling (article 4.2)

Profiling System that predicts the probability that an individual having a feature F_1 also has a feature F_2.

Profiling

In the GDPR, it is defined as any form of processing of personal data of a natural person that produces legal effects (e.g., signing a contract) or significantly affects it. It includes analyses and predictions related to work, economic situation, health, interests, reliability, location, ...

According to the European Data Protection Board, profiling is the process of classifying individuals or groups into categories based on their features.

Example (Cambridge Analytica scandal)**.** Case where data of US voters was used to identify undecided voters:

1. US voters were invited to take a personality/political test that was supposed to be for academic research. Participants were also required to provide access to their Facebook page in order to get a money reward for the survey.
2. Cambridge Analytica collected the participants' data on Facebook, but also accessed data of their friends.
3. The data of the participants was used to build a training set where Facebook content is used as features and questionnaire answers as the target. The model built upon this data was then used for predicting the profile of their friends.

4. The final model was used to identify voters that were more likely to change their voting behavior if targeted with personalized ads.

1.4.1 Surveillance

Industrial capitalism Economic system where entities that are not originally meant for the market are also considered as products. This includes labor, real estate, and money. Industrial capitalism

Surveillance capitalism Considers human experience and behavior also as a marketable entity. Surveillance capitalism

Remark. Labor, real estate, and money are mostly subject to law. However, exploitation of human experience is less regulated.

Surveillance state System where the government uses surveillance, data collection, and analysis to identity problems, govern population, and deliver social services. Surveillance state

Example (Chinese social credit system)**.** System that collects data and assigns a score to citizens. The overall score governs the access to services and social opportunities.

1.4.2 Differential inference

Differential inference Make different predictions depending on the input features. Differential inference

In the context of profiling, it leads individuals with different features to a different treatment.

Example (ML in healthcare)**.** Using machine learning to predict health issues provides benefits to all data subjects. Processing data in this way is legitimate as long as appropriate measures are taken to mitigate privacy and data violation, and the overall risks are proportionate to the benefits.

Example (ML in insurance/recruiting)**.** Using machine learning with health data for recruiting or determining insurance policies would worsen the situation of who is already disadvantaged. Also, having the ability of distinguishing applicants creates a competitive advantage that leads to collect as much personal data as possible.

Distributive justice Theory based on the allocation of resources aiming for social justice. Distributive justice

Example (Price differentiation)**.** Differentiate prices based on the economic availability of the buyer allows for a generally higher accessibility of goods.
However, if certain protected features are used to determine the price instead, it would result in unfairness and exclusion.

1.4.3 Discrimination

There are two main opinions on AI systems:

- AI can avoid fallacies of human psychology (e.g., overconfidence, loss aversion, anchoring, confirmation bias, ...).
- AI can make mistakes and discriminate.

  Direct discrimination/Disparate treatment When the AI system bases its prediction on protected features.

  Indirect discrimination/Disparate impact The AI system has a disproportional impact on a protected group without a reason.

Remark. AI systems trained on a supervised dataset might:

- Reproduce past human judgement.
- Correlate input features to (not provided) protected features (e.g., ethnicity could be inferred based on the postal code).
- Discriminate groups with common features (e.g., the number of working hours of women are historically lower than men).
- Lead to unfairness if the data does not reflect the statistical composition of the population.

1.5 Consent (article 4.11)

Consent

Consent Agreement of the data subject that allows to process its personal data. Consent should be:

Freely given The data subject have the choice to give consent for profiling

Remark. A common practice is the "take-or-leave" approach, which is illegal.

Remark. Showing the deny button in a less noticeable style is also not considered freely given.

Remark. Making the user pay the service if it does not consent to profiling is lawful.

Specific A single consent should be related to personal data used for a specific purpose and compatible ones.

Remark. A single checkbox for lots of purposes is illegal.

Informed The data subject should be clearly informed of what it is consenting to.

Remark. In practice, privacy policies are very vague.

Unambiguously provided Consent should be explicitly provided by the data subject through a statement of affirmative action.

Remark. An illegal practice in many privacy policies is to state that there can be changes and continuing using the service implies an implicit acceptance of the new terms.

1.5.1 Conditions for consent (article 7)

Conditions for consent

Some requirements for consent are:

- The controller must be able to demonstrate that the data subject has provided its consent.
- If consent for data processing is provided in written form alongside other matters, it should be clearly distinguishable.
- The data subject have the right to easily withdraw its consent at any time. The withdrawal does not affect previously processed data.
- To consider consent for profiling freely given, it should be assessed whether the performance of a contract is conditional on consenting the processing of personal data (i.e., the "take-or-leave" approach is illegal).

- Consent is by default considered not freely given in case of imbalance between the data subject and the controller, unless it can be proved that there were no risks if the data subject refused to consent.

1.6 Data subjects' rights

1.6.1 Controllers' information duties (articles 13-14)

Controllers' information duties

When personal data is collected, the controller should provide the data subject with the following information:

- The identity of the controller, its representative (when applicable), and its contact details should be available.
- Contact details of the data officer (referee of the company that ensures that the GDPR is respected) should be available.
- Purposes and legal basis of the processing.
- Categories of data collected.
- Recipients or categories of recipients.
- Period of time or the criteria to determine how long the data is stored.
- Existence of the rights to access, rectify, transfer, and erase data.
- Possibility to lodge a complaint with supervisory authorities.
- Source where the data originate (e.g., directly, from another account).
- Existence of automated decision-making systems based on profiling.

Moreover, in case of automated decision-making, the following information should be ideally provided:

- Input data that the system takes and how different data affects the outcome.
- The target value the system is meant to compute.
- The possible consequences of the automated decision.
- The overall purpose of the system.

1.6.2 Right to access (article 15)

Right to access

Data subjects have the right to have confirmation from the controller on whether their data has been processed and access both input and inferred personal data.
This right is limited if it affects the rights or freedoms of others.

1.6.3 Right to rectification

Right to rectification

Data subjects, depending on the case, have the right to rectify their personal data:

- In the public sector, there should be procedures when allowed.
- In the private sector, right to rectification should be balanced with the respect for autonomy of private assessments and decisions.

Generally, data can be rectified when:

- The correctness can be objectively determined.
- The inferred data is probabilistic and there was either a mistake during inference or additional data can be provided to change the outcome.

1.6.4 Right to erasure (article 17)

Right to erasure

Data subjects have the right to have their own personal data erased without delay from the controller when:

- The data is no longer necessary for the purpose it was collected for.

  Example. An e-shop cannot delete the address until the order is arrived.
- The data subject has withdrawn its consent, unless there are other legal basis.
- The data subject objects to the processing and there are no overriding legitimate interests.

  Example. After cancelling from a mailing list, the email stored by the processors should be deleted.
- The data has been unlawfully processed.
- The data have to be erased for legal obligations.

  Example. After a period of time, archived exams have to be erased.

Remark. When the controller has shared personal data with third parties and erasure of that data is requested, it has to inform the other parties.

Also, the right to erasure does not apply if:

- It is to exercise the right of freedom of expression and information.
- Compliance with legal obligations.
- For public interest in public healthcare, scientific or historical research, statistical purposes (if anonymized).
- For legal and defense claims.

1.6.5 Right to portability (article 19)

Right to portability

Data subjects, when personal data has been collected through consent, have the right to receive their data from the controller in a machine-readable format that can be transferred to another controller.

1.6.6 Right to object (article 21)

Right to object

Data subjects have the right to request the termination of the processing of their data when all the following conditions are met:

- The data subject has reasons to withdraw.
- The reason for processing is for public interest or legitimate interests.
- The controller cannot demonstrate legitimate interests for processing the data.

Remark. If processing is based on consent, this right does not apply as the data subject can simply withdraw its consent.

Remark. Right to object also applies to:

- Profiling,
- Direct marketing (in any situation),
- Research and statistical purposes, unless it is done in the public interest.

1.6.7 Rights with automated decision-making (article 22)

Rights with automated decision-making

The data subject has the right to not have decisions based only on automated profiling if it produces legal effects or significant effects. Moreover, it should at least have the rights to:

- Obtain human intervention.
- Express its own point of view.
- Challenge the decision.

Remark. A negated right is an obligation.

Exceptions are applied when:

- Data is needed to enter or perform the contract.

  Example. It is allowed to use automated systems to process a high number of job applications.

- Authorization is given by the authorities.
- Explicit consent is given.

1.6.8 Explainability in the GDPR (article 22, recital 71)

Explainability in the GDPR

It is not clear whether the GDPR considers the right to explanation an obligation of the controller. Due to the fact that recital 71 mentions the right to an explanation while article 22 does not, there are two possible interpretations:

- Explanation is not legally enforceable, but it is recommended.
- As article 22 contains the qualifier "at least", explanation is legally required when possible.

Remark. Development of explanation techniques can be split into two main areas:

Computer science Provide understandable models from black-box systems. Techniques

in this field are usually intended for other experts and assume full access to the model. Example of methods are:

Model explanation Model the black-box system using an interpretable model.

Model inspection Analyze properties of the black-box model on different inputs.

Outcome explanation Extract the reason that lead to a particular outcome.

Social science Provide explanations understandable for the end-user. Example of approaches are:

Contrastive explanation Specify which input values made the difference (related to model inspection).

Selective explanation Focus on factors that are more relevant to human judgement.

Causal explanation Focus on the causes rather than statistical correlations.

Social explanation Tailor the explanation based on the individual's comprehension capability.

1.7 Risk-based data protection

Risk-based legislation Measures with the goal of actively preventing risks.

Risk-based legislation

1.7.1 Data protection by design and by default (article 25)

The controller must, both while designing and deploying the processing system, implement technical and organizational measures to respect data protection principles. It must also ensure that only the necessary data is processed for each purpose.

Data protection by design and by default

1.7.2 Impact assessment (articles 35-36)

Controllers must preventively perform impact assessment to processing systems that are likely to have high risks in terms of rights and freedoms of the data subjects. If the risk is high, the controller must consult the supervisory authority (i.e., national data protection authority) which will provide its written advice.

Data protection impact assessment

1.7.3 Data protection officers (article 37)

Controllers must appoint a data protection officer to ensure compliance with the GDPR if processing requires continuous monitoring on data subjects, involves large scale sensitive data, or concerns criminal convictions.

Data protection officers

2 CLAUDETTE

CLAUDETTE Clause detector (CLAUDETTE) is a system to classify clauses in terms of services or privacy policies as: CLAUDETTE

- Clearly fair,
- Potentially unfair,
- Clearly unfair.

Unfair contractual term (directive 93/13 art 3.1) A contractual term, that was not individually negotiated, is considered unfair if it causes a significant unbalance in the parties' rights and obligations. Unfair contractual term

2.1 Unfairness categories

Consent by using clause A clause is classified as: Consent by using clause

- potentially unfair, if it states that the consumer accepts the terms of service by simply using the service.

Privacy included A clause is classified as: Privacy included

- potentially unfair, if it states that the consumer consents to the privacy policy by simply using the service.

Unilateral change A clause is classified as: Unilateral change

- potentially unfair, if the provider can unilaterally modify the terms of service or the service.

Jurisdiction clause A clause is classified as: Jurisdiction clause

- clearly fair, if consumers have the right to raise disputes in their place of residence.
- clearly unfair, if it only allows judicial proceedings in a different city or country.

Choice of law A clause is classified as: Choice of law

- clearly fair, if the law of the consumer's country of residence is applied in case of disputes.
- potentially unfair, in any other case.

Arbitration clause A clause is classified as: Arbitration clause

- clearly fair, if arbitration is optional before going to court.
- clearly unfair, if arbitration should take place in a coutry different from the consumer's residence or should be based on the arbiter's discretion (and not by law).

- POTENTIALLY UNFAIR, in any other case.

Limitation of liability A clause is classified as: Limitation of liability

- CLEARLY FAIR, if the provider may be liable.
- POTENTIALLY UNFAIR, if the provider is never liable unless obliged by law.
- CLEARLY UNFAIR, if the provider is never liable (intentional damage included).

Unilateral termination A clause is classified as: Unilateral termination

- POTENTIALLY UNFAIR, if the provider has the right to suspend or terminate the service and the reasons are specified.
- CLEARLY UNFAIR, if the provider can suspend or terminate the service for any reason.

Content removal A clause is classified as: Content removal

- POTENTIALLY UNFAIR, if the provider can delete or modify the user's content and the reasons are specified.
- CLEARLY UNFAIR, if the provider can delete or modify the user's content for any reason and without notice.

2.2 Methodology

Training data Manually annotated terms of service.