Ethics in Artificial Intelligence (Module 2)

Last update: 05 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 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).

Personal data

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

Processing

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

Controller

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.

1.1.3 Principles relating to processing of personal data (article 5)

Processing personal data should be done respecting the following principles:

- Lawfulness, fairness, and transparency.
- Purpose limitation.
- Data minimization.
- Data accuracy.
- Storage limitation.
- Integrity and confidentiality.
- Accountability.

1.2 Lawfulness of processing (article 6)

Processing of personal data is lawful if at least one of the following conditions apply:

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

- 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.

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

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 access Data subjects have the right to access both input and inferred personal data.

Right to access

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

Right to rectification

- 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.

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.

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

Freely given The data subject have the choice to give consent or use another alternative (e.g., pay the service).

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

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

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.

Conditions for consent (article 7) Some requirements for consent are:

Conditions for consent

- 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 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).