Ethics in Artificial Intelligence (Module 2)

Last update: 03 May 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.

1.2 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)

Storage limitation

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

1.3 Personal data (article 4.1)

1.3.1 Identifiability

Identifiability

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

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

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

Anonymization

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

> **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

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

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

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

Unfair contractual term

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.

2.1 Unfairness categories

Consent by using clause

Consent by using clause A clause is classified as:

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

Privacy included

Privacy included A clause is classified as:

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

Unilateral change

Unilateral change A clause is classified as:

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

Jurisdiction clause

Jurisdiction clause A clause is classified as:

- 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

Choice of law A clause is classified as:

- 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

Arbitration clause A clause is classified as:

- 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.

Tasks Two tasks are solved:

Detection Binary classification problem aimed at determining whether a sentence contains a potentially unfair clause.

Sentence classification Classification problem of determining the category of the unfair clause.

Experimental setup Leave-one-out where one document is used as test set and the remaining as train ($\frac{4}{5}$) and validation ($\frac{1}{5}$) set.

Metrics Precision, recall, F1.

2.2.1 Base clause classifier

Experimented methods were:

- Bag-of-words,
- Tree kernels,
- CNN,
- SVM,
- ...

2.2.2 Background knowledge injection

Memory-augmented neural network

Memory-augmented neural network Model that, given a query, retrieves some knowledge from the memory and combines them to produce the prediction.

In CLAUDETTE, the knowledge base is composed of all the possible rationales for which a clause can be unfair. The workflow is the following:

1. The clause is used to query the knowledge base using a similarity score and the most relevant rationale is extracted.
2. The rationale is combined with the query.
3. Repeat the extraction step until the similarity score is too low.
4. Make the prediction and provide the rationales used as explanation.

Example (Knowledge base for liability exclusion)**.** Rationales are divided into six class of clauses:

- Kind of damage,
- Standard of care,
- Cause,
- Causal link,
- Liability theory,
- Compensation amount.

2.2.3 Multilingualism

Training data Same terms of service of the original CLAUDETTE corpus selected according to the following criteria:

- The ToS is available in the target language,
- There is a correspondence in terms of version or publication date between the documents in the two languages,
- There are structure similarities between the documents in the two languages.

Approaches Different strategies have been experimented with:

Novel corpus for target language

Novel corpus for target language Retrain CLAUDETTE from scratch with newly annotated data in the target language.

Semi-automated creation of corpus through projection

Semi-automated creation of corpus through projection Method that works as follows:

1. Use machine translation to translate the annotated English document in the target language while projecting the unfair clauses.
2. Match the machine translated document with the original document in the target language and project the unfair clauses (through human annotation).
3. Train CLAUDETTE from scratch.

Training set translation

Training set translation Translate the original document to the target language and train CLAUDETTE from scratch.

Remark. This method does not require human annotation.

Machine translation of queries

Machine translation of queries Method that works as follows:

1. Translate the document from the target language to English.
2. Feed the translated document to CLAUDETTE.
3. Translate the English document back to the target language.

> **Remark.** This method does not require retraining.

2.3 CLAUDETTE and GDPR

CLAUDETTE for GDPR compliance To integrate CLAUDETTE as a tool to check GDPR compliance, three dimensions, each containing different categories (ranked with three levels of achievement), are checked:

Comprehensiveness of information

Comprehensiveness of information Whether the policy contains all the information required by articles 13 and 14 of the GDPR.

Categories of this dimension comprises:

- Contact information of the controller,
- Contact information of the data protection officer,
- Purpose and legal bases for processing,
- Category of personal data processed,
- . . .

Substantive compliance

Substantive compliance Whether the policy processes personal data complying with the GDPR.

Categories of this dimension comprises:

- Processing of sensitive data,
- Processing of children's data,
- Consent by using, take-or-leave,
- Transfer to third parties or countries,
- Policy change (e.g., if the data subject is notified),
- Licensing data,
- Advertising.

Clarity of expression

Clarity of expression Whether the policy is precise and understandable (i.e., transparent).

Categories of this dimension comprises:

- Conditional terms: the performance of an action is dependent on a variable trigger.

  > **Remark.** Typical language qualifiers to identify this category are: depending, as necessary, as appropriate, as needed, otherwise reasonably, sometimes, from time to time, . . .

  > **Example.** "*We also may share your information if we believe, in our sole discretion, that such disclosure is necessary . . .*"

- Generalization: terms to abstract practices with an unclear context.

> **Remark.** Typical language qualifiers to identify this category are: generally, mostly, widely, general, commonly, usually, normally, typically, largely, often, primarily, among other things, . . .

> **Example.** "*We typically or generally collect information . . . When you use an Application on a Device, we will collect and use information about you in generally similar ways and for similar purposes as when you use the TripAdvisor website.*"

- Modality: terms that ambiguously refer to the possibility of actions or events.

  > **Remark.** Typical language qualifiers to identify this category are: may, might, could, would, possible, possibly, . . .
  > Note that these qualifiers have two possible meanings: possibility and permission. This category only deals with possibility.

  > **Example.** "*We may use your personal data to develop new services.*"

- Non-specific numeric quantifiers: terms that are ambiguous in terms of actual measure.

  > **Remark.** Typical language qualifiers to identify this category are: certain, numerous, some, most, many, various, including (but not limited to), variety, . . .

  > **Example.** "*. . . we may collect a variety of information, including your name, mailing address, phone number, email address, . . .*"

2.4 LLMs and privacy policies

> **Remark.** The GDPR requires two competing properties for privacy policies:
>
> **Comprehensiveness** The policy should contain all the relevant information.
>
> **Comprehensibility** The policy should be easily understandable.

Comprehensive policy from LLMs Formulate privacy policies for comprehensiveness and let LLMs extract the relevant information.

A template for a comprehensive policy could include:

- Categories of personal data collected,
- Purpose each category of data is processed for,
- Legal basis for processing each category,
- Storage period or deletion criteria,
- Recipients or categories of recipients the data is shared with, their role, the purpose of sharing, and the legal basis.

Experimental setup The following questions were defined to assess a privacy policy:

1. What data does the company process about me?
2. For what purposes does the company use my email address?
3. Who does the company share my geolocation with?
4. What types of data are processed on the basis of consent, and for what purposes?

5. What data does the company share with Facebook?
6. Does the company share my data with insurers?
7. What categories of data does the company collect about me automatically?
8. How can I contact the company if I want to exercise my rights?
9. How long does the company keep my delivery address?

Three scenarios were considered:

- Human evaluation of the questions on existing privacy policies,
- LLMs to answer the questions on ideal mock policies (with human evaluation).
- LLMs to answer the questions on real policies (with human evaluation).

Results show that:

- LLMs have high performance on the mock policies.
- LLMs and humans struggle to answer the questions on real privacy policies.

3 Discrimination

Disparate treatment The outcome of an algorithm is based on protected features.

Disparate treatment

Disparate impact The outcome of an algorithm that uses neutral features is disproportionate against certain groups without an acceptable reason.

Disparate impact

3.1 Biased data

3.1.1 Historical bias

Historical bias System trained on intrinsically biased data will reproduce the same biased behavior.

Historical bias

> **Remark.** Data can be biased because it comes from past human judgement or by the hierarchies of the society (e.g., systems working on marginalized languages will most likely have lower performance compared to a widespread language).

Example (Amazon AI recruiting tool)**.** Tool that Amazon used in the past to review job applications. It was heavily biased towards male applicants and, even with the gender removed, it was able to infer it from the other features.

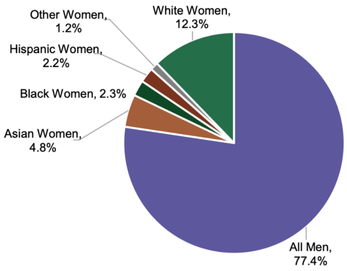
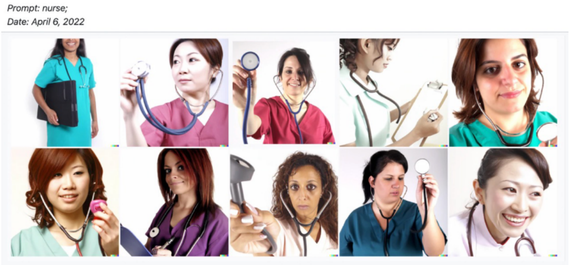


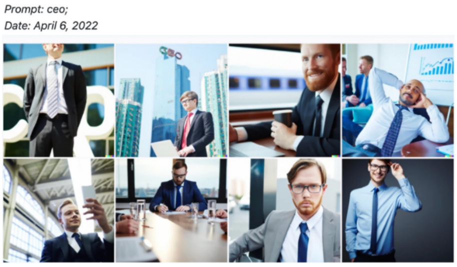
Figure 3.1: Tech companies workforce in the US

Example (UK AI visa and asylum system)**.** System used by the UK government to assess visa and asylum applications. It was found that:

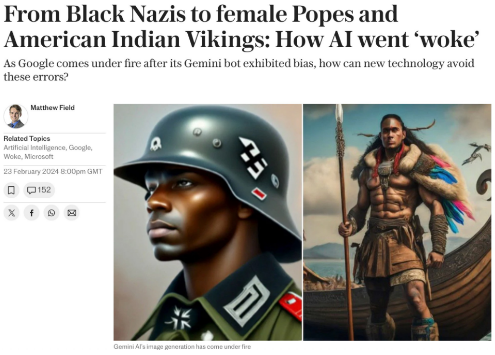
- The system ranked applications based on nationality.
- Applicants from certain countries were automatically flagged as high risk.

Example (Generative AI)**.** Prompting Stable Diffusion to generate the image of a `ceo` and a `nurse` highlights the gender and ethnicity bias of the training data.





Also, other systems (e.g., Gemini) included constraints to favor diversity resulting in unexpected results.



From Black Nazis to female Popes and American Indian Vikings: How AI went 'woke'

As Google comes under fire after its Gemini bot exhibited bias, how can new technology avoid these errors?

Matthew Field

Related Topics
Artificial Intelligence, Google, Woke, Microsoft

23 February 2024 8:00pm GMT

152

Gemini AI's image generation has come under fire

In the context of language models, some systems implement a refusal mechanism to prevent a biased response. However:

- Using a different prompt on the same topic might bypass the filter.
- Refusal might be applied unequally depending on demographics or domain.

3.1.2 Proxy variables

Proxy variable Neutral feature that is connected to a protected one resulting in a disparate impact on a certain group.

Proxy variable

Example (Disaster relief allocation system)**.** A system that predicts which communities need assistance based on past insurance claims is biased as it is a proxy for socioeconomic conditions: low-income communities often have lower insurance coverage and therefore will be disadvantaged by this system.

Example (Optum healthcare algorithm)**.** System used in US hospitals to predict which patients would benefit from additional resources. It was trained on historical data and it was found out that it was using the past healthcare cost data as a proxy to assess medical needs.
Due to historical disparity in accessing healthcare, this would cause a disparate impact on minorities that were unable to afford healthcare.

Example (Hurricane Katrina and racial disparities)**.** Due to historical racial segregation, the neighborhoods of many US cities can be divided by ethnicity. When Hurricane Katrina hit New Orleans, it mainly damaged the side of the city mostly lived by low-income communities. However, the evacuation plans assumed the availability of private vehicles and shelters were mostly built in the wealthier areas. Also, federal aid arrived quicker for wealthier communities and many low-income residences were never rebuilt.
An AI system trained on these data would reproduce the same behavior using the area one lives as a proxy.

3.1.3 Biases embedded in predictors

Bias embedded in predictors A system that uses favorable features that only a certain group has.

Bias embedded in predictors

Example (House allocation in Los Angeles)**.** VI-SPDAT is a system used in Los Angeles to distribute housing resources to homeless. As it relied on self-reported information, it was favoring those with higher literacy levels.

3.1.4 Unbalanced samples

Unbalanced samples

Unbalanced samples The dataset does not reflect the statistical composition of the population.

Example. Due to the lack of data of certain groups, a system to predict diseases will be more inaccurate towards minorities.

3.2 Algorithm choice

3.2.1 Aggregation bias problem

Aggregation bias problem

Aggregation bias problem System that has good results overall but with poor performance for specific groups.

Example. A system to predict the distribution of humanitarian aid trained on past successful data can present aggregation bias due to geographical data as a large part of the training data will most likely come from well established urban areas.

3.2.2 Different base rates

Base rate/prior probability

Base rate/prior probability Proportion of samples belonging to a certain class.

Example (COMPAS system)**.** COMPAS is a system used by US courts to determine the risk of recidivism (high, medium, low).

Loomis case E. Loomis was a defendant that according to COMPAS had a high risk of recidivism and was sentenced to 6 years in prison. The decision was appealed by Loomis as COMPAS has the following issues:

- Its functioning is unknown,
- Its validity cannot be verified,
- It discriminates on gender and ethnicity,
- Statistical predictions violate the right to individualized decisions.

The Supreme Court of Wisconsin rejected the argument and stated that:

- Statistical algorithms do not violate the right to individualized decisions as they are used to enhance a judge's evaluation,
- Gender is necessary to achieve statistical accuracy,
- Judges should be informed about the possibility of racial discrimination by COMPAS.

ProPublica and Northpointe studies ProPublica, a non-profit organization, published a study on the accuracy and fairness of COMPAS by comparing the predicted recidivism rates of 11 757 defendants and the actual rates between 2013 and 2014. Results found out that:

- The overall accuracy is moderate-low (61.2%),
- Black defendants were more likely labeled with a high level of risk, leading to a higher probability of high risk misclassification (45% blacks vs 23% whites).
- White defendants were more likely labeled with a low level of risk, leading to a higher probability of low risk misclassification (48% blacks vs 28% whites).

Northpointe, the software house of COMPAS, stated that ProPublic made several statistical and technical errors as:

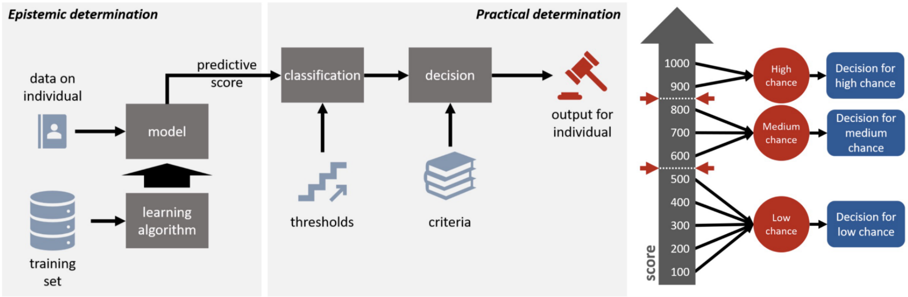
- The accuracy of COMPAS is higher that human judgement.
- The general recidivism risk scale is equally accurate for blacks and whites,
- COMPAS is compliant with the principle of fairness and does not implement racial discrimination.

Remark (Decision workflow)**.** A decision system can be represented in three steps:

1. Assign a predictive score (i.e., compute likelihood). In this step, unfairness can be caused by using protected features, biased data, a proxy, . . .

2. Classify the score based on some thresholds. In this step, unfairness can be caused by the choice of the threshold.

3. Make the decision. In this step, unfairness can be caused by how the value is used.



SAPMOC case SAPMOC is a toy example that predicts recidivism only based on whether the defendant has a previous criminal record. Assume that:

- 80% of previous offenders recidivate and the remaining do not.
- 20% of first time offenders recidivate and the remaining do not.
- The training data is composed of 3000 defendants divided into 1500 blues (1000 previous offenders) and 1500 greens (500 previous offenders).

Therefore, the real aggregated outcomes are:

	Has record	**No record**
Recidivism	1200 (80%)	300 (20%)
No recidivism	300 (20%)	1200 (80%)

Assume that SAPMOC's predictions are:

Table 3.1: Aggregated predictions

	Has record	**No record**
Recidivism	1500	0
No recidivism	0	1500

Table 3.2: Group-wise confusion matrix

	Pos.	TP	FP	**Neg.**	TN	FN
Blues	1000	800	200	500	400	100
Greens	500	400	100	1000	800	200

The base rates are then computed as:

	Base rate$_{\textbf{pos}}$ $\frac{\texttt{TP}+\texttt{FN}}{\texttt{TP}+\texttt{FN}+\texttt{FP}+\texttt{TN}}$	**Base rate**$_{\textbf{neg}}$ $\frac{\texttt{TN}+\texttt{FP}}{\texttt{TP}+\texttt{FN}+\texttt{FP}+\texttt{TN}}$
Blues	$\frac{900}{1500} = 60\%$	$\frac{600}{1500} = 40\%$
Greens	$\frac{600}{1500} = 40\%$	$\frac{900}{1500} = 60\%$

Note that the overall accuracy is the same for each group:

	Accuracy $\frac{\texttt{TP+TN}}{\texttt{TP+FN+FP+TN}}$
Blues	80%
Greens	80%

Fairness criteria The main fairness criteria are the following:

Statistical parity Each group should have an equal proportion of positive and negative predictions.

Statistical parity

Example (SAPMOC)**.** SAPMOC does not satisfy statistical parity:

	Predicted pos. $\frac{\texttt{TP+FP}}{\texttt{TP+FN+FP+TN}}$	**Predicted neg.** $\frac{\texttt{TN+TN}}{\texttt{FN+FN+FP+TN}}$
Blues	67%	33%
Greens	33%	67%

Equality of opportunity/true positive rate The members sharing the same features between different groups should be treated equally (i.e., same recall).

Equality of opportunity/true positive rate

Example (SAPMOC)**.** SAPMOC does not satisfy equality of opportunity:

	Recall pos. $\frac{\texttt{TP}}{\texttt{TP+FN}}$	**Recall neg.** $\frac{\texttt{TN}}{\texttt{TN+FP}}$
Blues	89%	67%
Greens	67%	89%

Calibration The proportion of correct predictions should be equal for each class within each group (i.e., same precision).

Calibration

Example (SAPMOC)**.** SAPMOC satisfies calibration:

	Precision pos. $\frac{\texttt{TP}}{\texttt{TP+FP}}$	**Precision neg.** $\frac{\texttt{TN}}{\texttt{TN+FN}}$
Blues	80%	80%
Greens	80%	80%

Conditional use error/false rate The proportion of incorrect predictions should be equal for each class within each group.

Conditional use error/false rate

Example (SAPMOC)**.** SAPMOC satisfies conditional use error/:

	False rate pos. $\frac{\texttt{FP}}{\texttt{TP+FP}}$	**False rate neg.** $\frac{\texttt{FN}}{\texttt{TN+FN}}$
Blues	20%	20%
Greens	20%	20%

Treatment equality The error ratio of positive and negative predictions should be equal across all groups.

Treatment equality

Example (SAPMOC)**.** SAPMOC does not satisfy treatment equality:

	Error pos. $\frac{\texttt{FP}}{\texttt{FN}}$	**Error neg.** $\frac{\texttt{FN}}{\texttt{FP}}$
Blues	200%	50%
Greens	50%	200%

Remark. There are many other fairness criteria that are correlated to those above.

Remark. There is a conflict between individual and group fairness so that not all criteria can be satisfied at once.

Handling different base rates

Do nothing Accept that different groups are actually associated to different probabilities.

Do nothing

Modify the threshold for everyone Raise (or decrease) the threshold to diminish the favorable classification for everyone (affecting more the groups with a higher base rate).

Modify the threshold for everyone

Change the decision for everyone Adopt alternative measures based on the classification results or use different thresholds depending on the group.

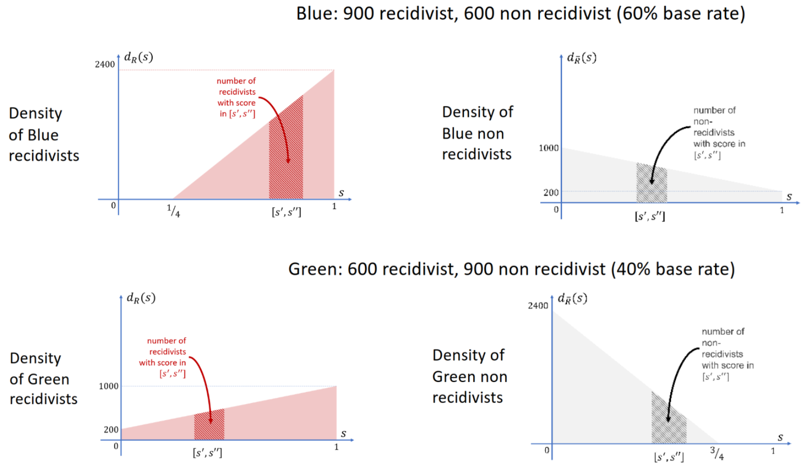
Change the decision for everyone

> **Remark.** Using different thresholds might still lead to discrimination. It makes sense in cases that require an affirmative action to increase diversity.

Example (SAPMOC II)**.** SAPMOC extended to multiple features and an output in $[0, 1]$. It is possible to represent the relationship between the output score and the likelihood of recidivism as densities:

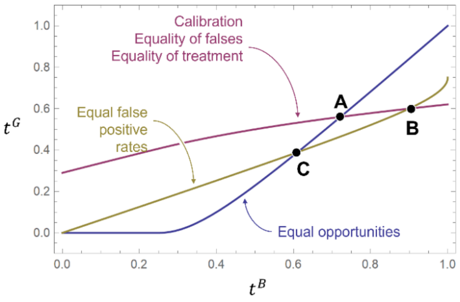
Recidivism density Function of the score such that the area under the curve between $[s', s'']$ is the number of recidivists associated to a score in that interval.

Non-recidivism density Function of the score such that the area under the curve between $[s', s'']$ is the number of non-recidivists associated to a score in that interval.



If the same threshold is applied for both groups, SAPMOC II respects the same fairness criteria of SAPMOC.

Theorem 3.2.1. With different base rates, it is impossible to achieve all fairness criteria through thresholding.


Calibration
Equality of falses
Equality of treatment
Equal false positive rates
Equal opportunities
A
B
C
t^G
t^B