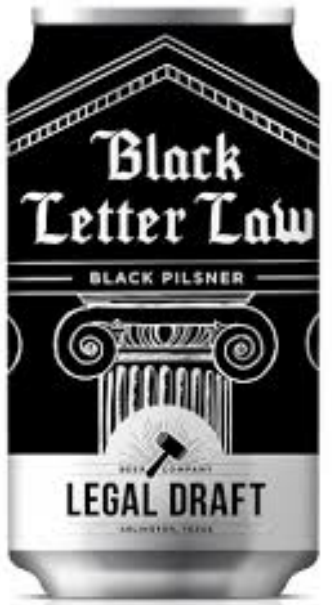
The right to explanation in the GDPR

Francesca Lagioia

Giovanni Sartor



AI and the data protection principles

- AI & Big Data challenge key data protection principles:
  - Article 5(1)(a) GDPR: Fairness, transparency (correttezza, trasparenza)
  - Article 5(1)(b) GDPR: Purpose limitation (limitazione della finalità)
  - Article 5(1)(c) GDPR: Data minimisation (minimizzazione dei dati)
  - Article 5(1)(d) GDPR: Accuracy (esattezza)
  - Article 5(1)(e) GDPR: Storage limitation (limitazione della conservazione)

Article 5(1)(a) GDPR: transparency

- The idea of **transparency** is specified in Recital 58, which focuses on conciseness, accessibility and understandability.
  - Il principio della trasparenza impone che le informazioni destinate al pubblico o all'interessato siano concise, facilmente accessibili e di facile comprensione e che sia usato un linguaggio semplice e chiaro, oltre che, se del caso, una visualizzazione. .

- Transparency in GDPR vs transparent and explainable AI.
  - providing sufficient information to lay people, relatively to issues that are relevant to them vs building a “scientific” model of the functioning of an AI system

Article 5(1)(a) GDPR: Informational Fairness

- **“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 existence of profiling and the consequences of such profiling.
- Informational fairness is also linked to **accountability (responsabilizzazione)**, since it presumes that the information to be provided makes it possible to check for compliance.

Substantive Fairness

- 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, which may be summarised by referring to the aforementioned **standards of acceptability, relevance and reliability:**

  - **use appropriate mathematical or statistical procedures** for the profiling,
  - **implement technical and organisational measures,** appropriate to ensure in particular, that .. inaccuracies in personal data are corrected and
  - **secure personal data** in a manner that takes account of the potential risks … and that prevents, inter alia, discriminatory effects

AI and transparency

- The issue of transparency can come up at two points in time
  - when a data subject's information is inputted in an information system that includes AI algorithms (ex-ante transparency), or
  - after the system's algorithmic model has been applied to the data subject, to deliver specific outcomes concerning his or her (ex-post transparency).

Information on automated decision making

The controller has the obligation to provide (Article 13(2)(f) and 14(2)(g) GDPR):

(a) information on “the existence of automated decision-making, including profiling, referred to in Article 22(1)” and

(b) “at least in those cases meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.”

A right to ex post explanation?

According to Recital (71), the safeguards to be provided to data subjects in case of automated decisions include all of the following:

- specific information
- the right to obtain human intervention,
- the right to express his or her point of view,
- **the right to obtain an explanation** of the decision reached after such assessment
- the right to challenge the decision.

According to Article 22 the suitable safeguards to be provided include “at least”

- the right to obtain human intervention,
- the right to express his or her point of view,
- the right to challenge the decision.

Information on automated decision making

- Computer scientists have focused on the technological possibility of providing understandable models of opaque AI systems (and, in particular, of deep neural networks), i.e., models of the functioning of such systems that can be mastered by human experts. For instance, the following kinds of explanations are at the core of current research on explainable AI:

  - **Model explanation**, i.e., the global explanation of an opaque AI system through an interpretable and transparent model that fully captures the logic of the opaque system.

  This would be obtained for instance, if a decision tree or a set of rules was provided, whose activation exactly (or almost exactly) reproduces the functioning of a neural network.

  - **Model inspection**, i.e., a representation that makes it possible to understanding of some specific properties of an opaque model or of its predictions.

  It may concern the patterns of activation in the system's neural networks, or the system's sensitivity to changes in its input factors (e.g. how a change in the applicant's revenue or age makes a difference in the grant of a loan application).

  - **Outcome explanation**, i.e., an account of the outcome of an opaque AI in a particular instance.

  For instance, a special decision concerning an individual can be explained by listing the choices that lead to that conclusions in a decision tree (e.g., the loan was denied because of the applicant's income fell below a certain threshold)

- The explanatory techniques and models developed within computer science are intended for technological experts and assume ample access to the system being explained.

Information on automated decision making

Social scientists have focused on the objective of making explanations accessible to lay people, thus addressing the communicative and dialectical dimensions of explanations. For instance, it has been argued that the following approaches are needed (Miller 2019, Mittelstadt and Wachter 2019).

- **Contrastive explanation**: specifying what input values made a difference, determining the adoption of a certain decision (e.g., refusing a loan) rather than possible alternatives (granting the loan);

- **Selective explanation**: focusing on those factors that are most relevant according to human judgement;

- **Causal explanation**: focusing on causes, rather than on merely statistical correlations (e.g., a refusal of a loan can be causally explained by the financial situation of the applicant, not by the kind of Facebook activity that is common for unreliable borrowers);

- **Social explanation**: adopting an interactive and conversational approach in which information is tailored according to the recipient's beliefs and comprehension capacities.

Information on automated decision making

- Ex-ante the user should ideally be provided with the following information:

- The input data that the system takes into consideration (e.g., for a loan application, the applicant's income, gender, assets, job, etc.), and whether different data items are favouring or rather disfavouring the outcome that the applicant hopes for;
- The target values that the system is meant to compute (e.g., a level of creditworthiness, and possibly the threshold to be reached in order for the loan to be approved);
- The envisaged consequence of the automated assessment/decision (e.g., the approval or denial of the loan application).
- It may also be useful to specify what are the overall purposes that the system is aimed to achieve

Is there a right to individual explanation?

Thus, two items are missing in article 22 relative to Recital (71): the provision of “specific information” and the right to *obtain an explanation of the decision reached after such assessment”*.

The second omission in particular raises the issue of whether controllers are really required by law to provide an individualised explanation

A right to explanation? Two possible interpretations

- According to the first interpretation,
  - providing individualised explanation would only be a good practice, and not a legally enforceable requirement.
- According to the second interpretation,
  - there is an enforceable legal obligation to provide individual explanation, unless this is impossible or too burdensome (proportionality assessment).

Summary: provisions on explanation

Main references in the GDPR:

- **Article 13 and 14** (on the right to information) and **Article 15** (on the right to access): the controller should provide information on “the existence of automated decision-making, including profiling, referred to in Article 22(1)” and “meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject”.
- **Article 22**: the data subject has at least the right to obtain human intervention, the right to express his or her point of view, and the right to challenge the decision.
- **Recital (71),** the data subject should also have the right to obtain an explanation of the decision reached after the assessment of his or her circumstances.

What rights to information and explanation?

Based on this set of norms, the obligation to provide information to the profiled data subject can take very different content:

1. information on the existence of profiling, i.e., on the fact that the data subject will be profiled or is already being profiled;
2. general information on the purposes of the profiling and decision making;
3. general information on the kind of approach and technology that is adopted;
4. general information on what inputs factors (predictors) and outcomes (targets/predictions), of what categories are being considered;
5. general information on the relative importance of such input factors in determining the outcomes;
6. specific information on what data have been collected about the data subject and used for profiling him or her;
7. specific information on what values for the features of the data subject determined the outcome concerning him or her;
8. specific information on what data have been inferred about the data subject;
9. specific information on the inference process through which certain values for the features of the data subject have determined a certain outcome concerning him or her.

(1-5 ex ante; 6 – 9 ex post)

Conclusion?

- Given the variety of ways in which automated decision-making can take place, it is hard to specify in precise and general terms what information should be provided. What information the controller may be reasonably required to deliver will indeed depend on the importance of the decision, on the space of discretion that is being used, and on technological feasibility.