Machine learning and profiling in the PNR system

Janneke Gerards

2023-05-08T15:02:13

The Passenger Name Record (PNR) Directive, adopted in 2016, has led to significant debate among lawyers. Several preliminary references have been made to enquire into its interpretation and validity. In 2022, one of these led to the landmark judgment of the *CJEU* in *Ligue des droits humains*. As this series of blogposts on the PNR judgment show, this is a rich judgment with many very interesting elements, among which are the CJEU's considerations related to non-discrimination. Indeed, automated processing of personal data, which is what PNR data are, can lead to forms of profiling, in the sense that certain individuals or groups of people are more likely to be excluded based on the transfer of their data than others. If those people are also found to have certain characteristics, such as a particular ethnic or national origin or religion, they may be directly or indirectly disadvantaged, in violation of the prohibition of discrimination. In its judgment, the CJEU extensively discusses these discrimination risks, and it set a number of conditions to prevent them. Unfortunately, as the present post aims to further explain, not all of its considerations are perfectly clear and some of the solutions the CJEU proposes are not entirely satisfactory.

Safeguards against discrimination in the PNR Directive

As such, the PNR Directive contains several safeguards against discrimination. According to Article 6(2) of the PNR Directive, a Passenger Information Unit (PIU; the national police units in charge of receiving, processing, and further sharing of the passenger data), may only process PNR data for a limited number of purposes. Most relevant when it comes to non-discrimination is the (a) ground. This relates to the assessment whether certain passengers should be subject to further examination because they might be involved in a terrorist offence or serious crime. For the purpose of that assessment, the competent PIU may firstly compare the PNR data with, for example, databases containing information related to wanted or reported persons (Art. 6(3)(a) of the PNR Directive). If that information leads to a match, the PIU may transmit the data to the competent authorities of a Member State, which may take action on that basis. In addition, Article 6(3)(b) allows a PIU to process PNR data "against pre-determined criteria". The assessment of the data according to these criteria must, according to paragraph 4, be carried out "in a non-discriminatory manner". Also, according to this paragraph, the predetermined criteria must be "targeted, proportionate and specific". They must be regularly reviewed and shall 'in no circumstances be based on a person's race, ethnic origin, religious, philosophical or political beliefs, trade union membership, health, sex life or sexual orientation". Insofar as relevant, Article 6(5) additionally stipulates that automatic processing operations to see whether certain passengers may be subject to further scrutiny should be checked on a case-by-case basis in a non-automated manner – that is, manually, by humans.

‘Pre-determined criteria’ and machine learning algorithms

As Advocate General Pitruzella has explained, setting such pre-determined criteria involves profiling: certain criteria are used to ‘predict’ which passengers might be involved in a terrorist crime or serious crime (para. 223). In a first evaluation of the PNR Directive, the European Commission gave the following definition of pre-determined criteria: “Pre-determined criteria, also known as targeting rules, are search criteria, based on the past and ongoing criminal investigations and intelligence, which allow to filter out passengers which correspond to certain abstract profiles, e.g. passenger travelling on certain routes commonly used for drug trafficking, who bought their ticket in the last moment and paid in cash, etc.” Accordingly, profiles are made based on individual characteristics, which, taken together, may reveal a risk for certain behaviour in the future. Unfortunately, however, the Directive does not explain how and on what grounds these targeting rules can be established, what data may be used to do so, or where that data may be sourced from.

In *Ligue des droits humains*, the CJEU concentrated on the use of these pre-determined criteria. In particular, it ruled that the requirement that criteria must be ‘pre-determined’ “[…] precludes the use of artificial intelligence technology in self-learning systems (‘machine learning’), capable of modifying without human intervention or review the assessment process […]” (para. 194). This consideration of the CJEU is rather ambiguous, but it seems to imply that the use of machine learning algorithms (hereafter: ‘ML algorithms’) to set pre-determined criteria can only be accepted if very strict conditions are met. However, to understand the importance of the PNR judgment for the use of ML algorithms, and to be able to deal with its ambiguity, it may be good to first briefly explain how such algorithms work. To put it very simply, an algorithm can be ‘taught’ to detect strong similarities and correlations in large datasets and to determine statistical relationships and probabilities – for instance, the probability that someone who is showing certain characteristics and behaviour is preparing a terrorist attack. This teaching and learning is done through a feedback system. For example, a programmer will give the algorithm positive feedback if it has correctly detected a person as setting a particular risk, as that person has been convicted of terrorism in the past. Conversely, the programmer will give negative feedback if the algorithm has wrongly identified someone as likely to be involved in terrorism. In this way, the algorithm’s pattern recognition is constantly being refined and improved. The more often this feedback process is repeated, the more accurately the algorithm’s predictions will become, until a point that it can be validated and be used in practice.

During the training process, the algorithm may be taught not to take into account any of the characteristics listed in Article 6(4). A well-trained algorithm is then unlikely to readily identify factors such as ‘origin of country x’ or ‘member of religious group y’ as criteria relevant to determining whether a passenger is likely to have terrorist or criminal intentions. Instead, the algorithm will look for other patterns in the plethora of information that may be available about people who have committed serious criminal offences or terrorism in the past. In doing so, an algorithm does not use the same causality-oriented logic as humans. Instead, an algorithm looks for statistically significant relationships (correlations) between certain factors. As a result, an algorithm may find, for example, that within a dataset there is a correlation between terrorist behaviour and seemingly illogically related factors such as late booking, searching the internet for information about planes, frequenting the toilet at the airport and sending messages in a certain language. Perhaps human beings might not easily think of the combination of such factors as relevant or causally related to terrorism. However, because it might be correlated to a risk of terrorism or serious crime, it still could provide a pre-determined criterion to start processing PNR data.

To many people, the method discussed above may seem very useful. However, the CJEU rightly identified a number of considerable drawbacks related to the deployment of ML algorithms and profiling, especially when viewed from the specific perspective of the prohibition of discrimination.

Discrimination-related problems of ML algorithms

A first problem is the so-called base rate fallacy. This means that, however carefully an algorithm is trained, it may still identify ‘false positives’ or ‘false negatives’. In other words, it can happen that the algorithm either wrongly designates a person as constituting a risk, or it misses a person who *would* present a risk. So that either means wrongly identifying someone as potentially suspicious, or overlooking an actual risk of terrorism or serious crime. Both are clearly problematic, but there is research pointing out that the PNR system causes an especially high risk of false positives.

Another problem is that a machine learning algorithm can only operate properly when it is trained on a good dataset. It must be relevant to the specific European context, it must contain enough data, it must be representative of the kind of information that is needed, and so on. It proves very difficult to compile or obtain those kinds of datasets and prevent them from reflecting discriminatory and stereotypical patterns in human thought and action. The well-known risk of ‘rubbish in is rubbish out’ then easily arises, in that deficiencies and discrimination in the data soon translate into inaccuracies and discrimination in the output of an algorithm. Moreover, if a dataset is not properly set up and prepared, it can severely disrupt the learning process. For instance, an algorithm may learn to recognise certain patterns as relevant on the basis of non-representative or incorrect data, when in reality those patterns turn out to be incorrect. In that case, the risk of false negatives and false positives increases even more. This risk is amplified if the algorithm continues to ‘improve’ itself in practice by recognising patterns independently in newly added (also coloured) data. This can lead to reinforcement of already existing forms of (institutional) discrimination. For PNR algorithms, the lack of good datasets is a well-known problem. The European Parliament's research office has pointed out that data obtained from airlines or under PNR agreements with third countries are very unreliable. As a result, there is a high risk of discrimination through the use of ML algorithms.

Third, the Court has observed that training an algorithm is similar to taking a snapshot. At a certain point in the training process, an algorithm is validated and found suitable to perform. But then, in practice, the datasets the algorithm works with are constantly changing. Information may continuously be added about suspected and convicted or acquitted people, or about their behaviour and habits. An algorithm that cannot adapt to such new data would quickly lose its relevance. Therefore, many ML algorithms are self-learning and can continue to update themselves by looking for useful correlations even in new data. That keeps the algorithm up-to-date, but it also makes it very easy to lose grip on how it works.

Finally, as noted above, risk assessments must not take account of protected personal characteristics, such as gender or ethnic origin. However, because an algorithm looks for correlations in a very fine-grained manner, it is difficult to fully prevent discrimination on these grounds. ML algorithms are easily able to establish correlations between seemingly harmless factors, such as times when someone is on their phone, distance from one's home to the airport or a preference for travelling by bus. Yet, in practice, even these factors can sometimes provide clues about someone's ethnicity, religious or political affiliation and may result in 'proxy discrimination'. It is difficult to eliminate this without making an algorithm completely ineffective. Moreover, it is far from easy to find out whether there is proxy discrimination, because it is usually not very clear exactly which constellations of factors an algorithm detects as a relevant pattern. Indeed, this is what the CJEU held against algorithmic profiling: "given the opacity which characterises the way in which artificial intelligence technology works, it might be impossible to understand the reason why a given program arrived at a positive match" (para. 195).

The CJEU's response

These problems explain why the CJEU has set such strict conditions for the use of ML algorithms in risk profiling in its judgment in the PNR case. Rather than having an algorithm do the work, the Court seems to prefer the pre-determined criteria either to be set or at least to be applied or checked by human beings, for example, by officials working at the PIUs. According to the Court, in processing the data, the PIU and the competent authorities "can inter alia take into consideration specific features in the factual conduct of persons when preparing and engaging in air travel which, following the findings of and experience acquired by the competent authorities, might suggest that the persons acting in that way may be involved in terrorist offences or serious crime" (para. 199). In addition, the pre-determined criteria must meet several other requirements: they cannot be directly or indirectly discriminatory; they must meet the requirements of purposefulness, specificity and proportionality; they must take account of both incriminating and exonerating elements; and the number of false positives must be limited as much as possible.

It is unclear how discrimination can be prevented by these requirements, which could explain why the judgment can appear to be rather confusing. In fact, it seems that the Court's suggestions in paragraph 199 mean that the role of a predictive algorithm is effectively taken over by humans, who would come up with risk assessments based on their own experiences and actual observations, and would correct the algorithmic output and predictions accordingly. This 'human in the loop' approach might seem attractive, but human beings, too, are prone to stereotyped thinking. The kind of patterns they think they can observe can be very stigmatising. Consequently, one risk of discrimination is simply replaced by another one. Moreover, it is far from obvious that human profiling would lead to fewer false positives and negatives, or would be more transparent than a human prediction. Hence, the Court's strong emphasis on 'human intervention or review' certainly is not a sufficient guarantee against discrimination.

Fortunately, the Court added several considerations on the procedural safeguards that should accompany the use of pre-determined criteria. Their use should be based on a coherent administrative practice in which the principle of non-discrimination is paramount (para. 205). Checks of algorithmic output should be based on clear, precise and objective monitoring criteria that can help determine whether a person is indeed potentially involved in terrorism or serious crime (para. 206). There should be accurate documentation of the processing to enable verification and internal control of its lawfulness and stakeholders should be able to understand the criteria and the programmes that work with them (para. 207). This would enable them to avail themselves of legal protection options. If those options are used, the courts should be able to take note of all the relevant criteria and check how the programmes work (para. 210). The same applies to national supervisory authorities, who should be able to check that there was no discrimination (para. 212).

Indeed, such procedural safeguards are highly useful and offer better protection against the discriminatory use of ML algorithms than the Court's suggestions as to human intervention. It might have been even better if the CJEU had also ruled that machine learning algorithms may only be deployed if sound non-discrimination safeguards are built into the programming and training processes and if mechanisms are put in place to detect and counteract discriminatory output. Nevertheless, to the extent the Court emphasises the need for multiple procedural safeguards when deploying algorithms, its ruling has considerable added value.