Douwe Korff
Emeritus Professor of International Law, London Metropolitan University
Associate, Oxford Martin School, University of Oxford

The limitations of and flaws in algorithmic/AI-based technologies*

There are three main problems with algorithmic/AI data mining-based detection of rare phenomena (such as terrorists and serious criminals in a general population) (all of which have been noted and warned about in the past, so I will draw on those earlier reports):

- The base-rate fallacy and its effect on false positives;
- Built-in biases; and
- Opacity and unchallengeability of decisions

i. The base-rate fallacy and its effects on false positives

At the Court hearing in the *PNR* case in July 2021, there were some exchanges on false positive rates. These focussed on the discrepancy in searches in Passenger Name Records for people who "may be" terrorists or serious criminals" between the number of initial "hits" and the number of cases passed on for "further examination" by "competent authorities" that resulted in a "match". As judge-rapporteur von Danwitz noted (and as also noted earlier in my opinion), according to the staff working paper there were initial "hits" in 0.59% of all one billion records checked, but only 0.11% of cases were passed on – and this was taken, by the judge-rapporteur and the other interlocuters, as suggesting a "false positive" rate of 81% (4,800,000/5,900,000) and a "true positive" rate of 19% (1,100,000/5,900,000). The judge-rapporteur wondered "whether the fallibility [*fiabilité*] of such a system [was] acceptable" and how that should be assessed:

> I would not suggest that PCR tests [to detect Covid-19 in the pandemic] are the appropriate point of reference, but given the [Covid-19 pandemic] crisis in which the find ourselves, I believe that a test with a 19% sensitivity would not be well received.

Mr von Danwitz referred to the report on PNR data prepared by me with Marie Georges in 2015 for the Consultative Committee of Convention No. 108 and to the notion of the "**base-rate fallacy**" discussed in it.[1] However, remarkably but not surprisingly, both these initial remarks and the subsequent exchanges at the hearing showed a serious lack of understanding of the statistical issues, even in relation to false-true positives and negatives, and the base-rate fallacy. This is not surprising because lawyers have long failed to properly "do" mathematics (*iudex non calculat*, as the Romans used to say). But in the digital age, mathematical and statistical ignorance is no longer acceptable. More specifically in relation to algorithm/AI-based profiling and data mining, lawyers and judges should know enough about basic mathematics and statistics to understand the issues and implications of those methodologies. Some crucial aspects of the practices are counter-intuitive – and are either not understood or deliberately obscured or even denied by proponents of such tools.

---

* This paper reproduces (with minor editorial changes) section 4.9(f), sub-section (fe), of my Opinion on Core Issues in the PNR CJEU Case, prepared at the request of the Fundamental Rights European Experts Group (FREE Group), November 2021 (hereafter "**my opinion**" or "**the opinion**"), available at: https://www.ianbrown.tech/wp-content/uploads/2021/12/KORFF-FREE-Paper-on-Core-Issues-in-the-PNR-Case.pdf The issues relating to the mining of PNR data to which that opinion, and thus this excerpt, relate are illustrative of the wider problems with AI-based data mining (which is why I am reproducing this section spearately).

[1] Douwe Korff & Marie Georges, Passenger Name Records, data mining & data protection: the need for strong safeguards, report prepared for the Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (T-PD) of the Council of Europe, June 2015, available at: https://rm.coe.int/16806a601b

Douwe Korff
Emeritus Professor of International Law, London Metropolitan University
Associate, Oxford Martin School, University of Oxford

In very simple layperson's terms, the base-rate fallacy means that if you are looking for very rare instances or phenomena in a very large dataset, you will inevitably obtain a very high percentage of false positives in particular – and this cannot be remedied by adding more or somehow "better" data: by adding hay to a haystack.

In more scientific terms, the concept of the base-rate fallacy denotes the fact that "*humans in general do not take the basic rate of incidence, the base-rate, into account when intuitively solving … problems [relating to] probability*".[2] Or as Robert Matthews put it: "*[humans tend to] neglect [the importance] of prior probabilities in judging the probability of events.*"[3]

Yet taking the base rate into account is actually crucial in such contexts – and will often lead to surprising outcomes, even for those that at a rational level understand the issue. To again quote Matthews:[4]

> [D]espite its potentially serious implications for many real-life issues, the base-rate error has yet to achieve wider recognition. This is certainly true in relation to the use of algorithmic/AI-based profiling and data mining in trying to identify "known" terrorists or serious criminals (who luckily form only a minute section of the general population), and even more so in any attempt to try and predict which previously "unknown" persons might be a terrorist or serious criminal.

The mathematics are not easy to understand for non-mathematicians – let alone for lawyers: see the box, below.

The mathematics behind the base-rate fallacy:

The base-rate fallacy is one of the cornerstones of Bayesian statistics, as it stems directly from Bayes' famous theorem (1):

$$P(A\backslash B) = \frac{P(A).P(B\backslash A)}{P(B)}$$

Expanding the probability *P(B)* for the set of all *n* possible, mutually exclusive outcomes *A* we arrive at equation (2):

$$P(B) = \sum_{i=1}^{n} P(Ai).P(B\backslash A)$$

Combining equations (1) and (2) we arrive at a generally more useful statement of Bayes' theorem:

$$P(A\backslash B) = \frac{P(A).P(B\backslash A)}{\sum_{i=1}^{n} P(Ai).P(B\backslash A)}$$

Source: Stefan Axelsson, The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection, 1999, p. 3.

---

[2] Stefan Axelsson, The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection, 1999, p. 4, available at: http://www.raid-symposium.org/raid99/PAPERS/Axelsson.pdf

[3] Robert Matthews. Base-rate errors and rain forecasts, Nature, Vol. 382(6594), p. 766, 29 August 1996, available at: https://www.nature.com/articles/382766a0.pdf

[4] *Idem*.

However, the implications can be well described through examples. Axelsson and Matthews provide the following two examples:[5]

Example 1: (Axelsson)

Suppose that your physician performs a test that is 99% accurate, i.e. when the test was administered to a test population all of which had the disease, 99% of the tests indicated disease, and likewise, when the test population was known to be 100% free of the disease, 99% of the test results were negative. Upon visiting your physician to learn of the results he tells you he has good news and bad news. The bad news is that indeed you tested positive for the disease. The good news however, is that out of the entire population the rate of incidence is only 1=10000, i.e. only 1 in 10000 people have this ailment. What, given the above information, is the probability of you having the disease?[6]

Leaving out the formulae in which the above data are applied, that draw on those in the box, above, the result is surprising and counter-intuitive:

Even though the test is 99% certain, your chance of actually having the disease (when tested positive) is only 1/100 (i.e., 1%), due to the fact that the population of healthy people is much larger than the population with the disease. In other words, the fact that the test is positive does not say much, in absolute terms, about your state of health.

Example 2: (Matthews)

The effect of the base-rate error can be explained with reference to a familiar (indeed, notorious) dilemma - that of how to respond to weather forecasts.

It seems obvious that decisions affected by the weather (going for a walk, for example) are best made by putting one's faith in the most accurate forecast available. Surprisingly, however, the base-rate effect can make this a sub-optimal approach.

The UK Meteorological Office's 24-hour forecasts of rain currently achieve around 83 per cent accuracy, while the probability of rain on the hourly timescale relevant to walks is around 0.08. The table below reveals the impact of the base-rate error in the interpretation of forecasts of rain.

---

[5] **Example 1** is taken from Stefan Axelsson, The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection, 1999, p. 3, available at: http://www.raid-symposium.org/raid99/PAPERS/Axelsson.pdf
The results are illustrated in a Venn diagram in Appendix A to Axelsson's paper, on p. 10. On the issue on which the paper focussed, intrusion detection, Axelsson concludes as follows:
"*[I]ntrusion detection in a realistic setting is perhaps harder than previously thought. This is due to the base-rate fallacy problem, and because of it, the factor limiting the performance of an intrusion detection system is not the ability to correctly identify behaviour as intrusive, but rather* ***its ability to suppress false alarms****. A very high standard, less than 1/100,000 per "event" given the stated set of circumstances, will have to be reached for the intrusion detection system to live up to these expectations, from an* ***effectiveness*** *standpoint. Much work still remains before it can be demonstrated that current IDS approaches will be able to live up to real world expectations of effectiveness.*"
(p. 8, original emphases were in italics.)

Example 2 is taken from Robert Matthews. Base-rate errors and rain forecasts, Nature, Vol. 382(6594), p. 766, 29 August 1996, available at: https://www.nature.com/articles/382766a0.pdf

[6] The reader is encouraged to make a quick "guesstimate" of the answer, at this point. [original footnote]

Table:

THE VARIOUS OUTCOMES OF FORECAST AND WEATHER OVER 1,000 1-HOUR WALKS			
	Rain	No rain	Sum
Forecast of rain	66	156	222
Forecast of no rain	14	764	778
Sum	80	920	1000

With forecast accuracies of 83 per cent, one might expect that a forecast of rain during the one-hour walk would be correct 83 per cent of the time. However, the hourly base-rate of rain in the United Kingdom is so low that forecasts of rain are more than twice as likely to be wrong as right: from the table, the probability of rain, given a forecast of rain - that is, *P*(rain/forecast of rain) – is 66/222=0.30, whereas *P*(no rain/forecast)= 156/222 =0.70.

This result suggests that those who ignore Meteorological Office forecasts may fare better than those who abide by them.

Thus, unless one is particularly concerned about getting wet, the base-rate effect makes disregard of forecasts of rain the optimal strategy.

Similar reasoning also reveals that, contrary to popular belief, always carrying an umbrella is a sub-optimal strategy unless one is morbidly afraid of getting wet. Indeed, [in the vast majority of cases] the base-rate effect makes even insouciant optimism a better strategy.

These examples clearly show the counter-intuitive nature of the base-rate fallacy.

The base-rate fallacy and the use of PNR Data: (EDRi/Epicenter)

In a write-up for *European Digital Rights* (EDRi), the Austrian digital rights organisation *Epicenter* linked the issue of the base-rate fallacy to the specific matter of the use of PNR data to "identify" terrorists and other serious criminals, with a very useful illustration, as follows:[7]

The Austrian implementation of the PNR Directive

In Austria, the Austrian Passenger Information Unit (PIU) has processed PNR since March 2019. On 9 July 2019, the Passenger Data central office (*Fluggastdatenzentralstelle*) issued a response to inquiries into PNR implementation in Austria. According to the document, from February 2019 to 14 May, 7,633,867 records had been transmitted to the PIU. On average, about 490 hits per day are reported, with an average of about 3,430 hits per week requiring further verification. According to the document, out of the 7,633,867 reported records, there were 51 confirmed matches and in 30 cases there was the intervention by staff at the airport concerned.

Impact on innocents

What this small show of success does not capture, however, is the damage inflicted on the thousands of innocent passengers who are wrongly flagged by the system and who can be

---

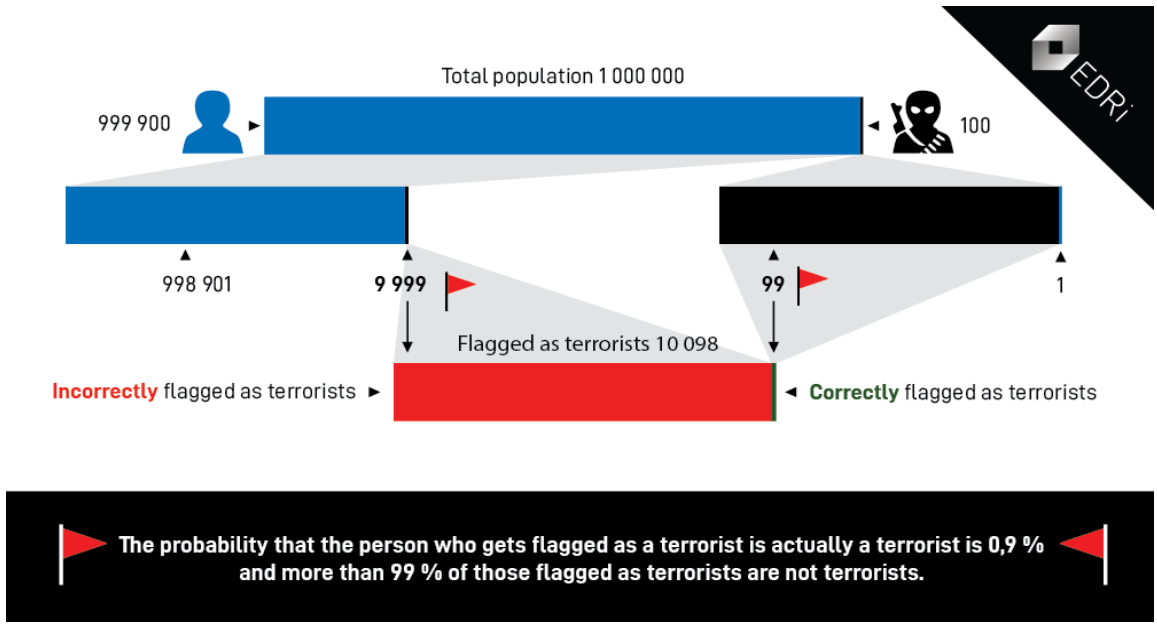
[7] Epicenter, *Why EU passenger surveillance fails its purpose*, 25 September 2019, available at: https://edri.org/our-work/why-eu-passenger-surveillance-fails-its-purpose/ (reproduced here with minor edits)

subjected to damaging police investigations or denied entry into destination countries without proper cause. Mass surveillance that seeks a small, select population is invasive, inefficient, and counter to fundamental rights. It subjects the majority of people to extreme security measures that are not only ineffective at catching terrorists and criminals, but that undermine privacy rights and can cause immense personal damage.

Why is this happening? The base-rate fallacy

Imagine a city with a population of 1,000,000 people implements surveillance measures to catch terrorists. This particular surveillance system has a failure rate of 1%, meaning that (1) [in relation to persons who are actually terrorists], the system will register [this] as a "hit" 99% of the time, and fail to do so 1% of the time and (2) [in relation to persons who are not terrorists], the system will not flag them 99% of the time, but register the person as a "hit" 1% of the time. What is the probability that a person flagged by this system is actually a terrorist?[8]

At first, it might look like there is a 99% chance of that person being a terrorist. Given the system's failure rate of 1%, this prediction seems to make sense. However, this is an example of incorrect intuitive reasoning because it fails to take into account the error rate of hit detection. This is based on the base-rate fallacy: The base rate fallacy is the tendency to ignore base rates – actual probabilities – in the presence of specific, individuating information. Rather than integrating general information and statistics with information about an individual case, the mind tends to ignore the former and focus on the latter. One type of base rate fallacy is the one suggested above called the false positive paradox, in which false positive tests are more probable than true positive tests. This result occurs when the population overall has a low incidence of a given condition and the true incidence rate of the condition is lower than the false positive rate. Deconstructing the false positive paradox shows that the true chance of this person being a terrorist is closer to 1% than to 99%:



[8] As in the Axelsson example, the reader should be encouraged to make a quick "guesstimate" of the answer, at this point (cf. footnote 5, above).

Note: The matter is in fact somewhat more complicated, in that there can be different accuracy rates for true positive rates (also known, in particular in medical research, as **sensitivity**) than for true negative rates (also known in such research as **specificity**).[9] Both can only be determined by reference with reference to a so-called "**Gold Standard**" – that is, the initially hypothetically assumed level of occurrence of the issue that is being looked for (which can then be refined on the basis of subsequently obtained real data). This where the base-rate fallacy (as well as another fallacy called "anchoring") come into play. As explained in relation to medical research (with a reference to Covid-19):[10]

> Interpretation of a test result depends not only on the characteristics of the test itself but also on the pre-test probability of disease. Clinicians use a heuristic (a learned mental short cut) called anchoring and adjusting to settle on a pre-test probability (called the anchor). They then adjust this probability based on additional information. This heuristic is a useful short cut but comes with the potential for bias. **When people fail to estimate the pre-test probability and only respond to a piece of new information, they commit a fallacy called base-rate neglect [or fallacy – DK]. Another fallacy called anchoring is failing adequately to adjust one's probability estimate, given the strength of new information.** Likelihood ratios can give a clinician an idea of how much to adjust their probability estimates. Clinicians intuitively use anchoring and adjusting thoughtfully to estimate pre- and post-test probabilities unconsciously in everyday clinical practice. However, faced with a new and unfamiliar disease such as covid-19, mental short cuts can be uncertain and unreliable and public narrative about the definitive nature of testing can skew perceptions.

Astonishingly, neither the report of the Commission on the review of the PNR Directive nor the accompanying staff working document mentions the base-rate fallacy. In fact, the reports do not even mention the base rate itself: the total number of individuals on whom PNR data are collected. As explained above, the effectiveness of the processing (filtering) of those data simply cannot be measured without taking that base rate into account. Worse***: failing to take that base rate into account is precisely what leads to the base-rate fallacy.***

So what is the base rate for the PNR checks? According to Eurostat, in 2019, 1,034 million people in the EU travelled by air; in 2018 the number was just under 1,000 million.[11] However, that number reflects the total number of flights undertaken by a person – so people who fly more than once in a year are recorded as two "person-flights"; one return flight also generates two PNRs. In fact, quite a few people will have travelled by air more than once, and many not at all. A 2014 survey suggested that 52% of the UK population had not flown at all in that year, while 15% of

---

[9] The terms were introduced by American biostatistician Jacob Yerushalmy in 1947, see: Jacob Yerushalmy, *Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques*, Public Health Reports (1896-1970), Vol. 62, No. 40, Tuberculosis Control Issue No. 20 (Oct. 3, 1947), pp. 1432-1449, available at: https://www.jstor.org/stable/4586294 ($)

[10] *Interpreting a covid-19 test result*, British Medical Journal, 2020; 369, 12 May 2020, available at: https://www.bmj.com/content/369/bmj.m1808.long (emphasis added)
The webpage contains a useful interactive "Covid-19 test calculator" on which we will draw below.

[11] According to Eurostat, in 2019, 1,034 million people in the EU travelled by air; in 2018 the number was just under 1,000 million. See:
https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Air_transport_statistics
The figures for 2020 are not in the normal range because of the Covid-19 pandemic.

that population had flown three or more times. On the other hand, one PNR can cover more than one person (in case of group bookings).

A very rough guess based on those statistics would be that on average the 1 billion people counted by Eurostat relate to 500 million distinct individuals (each taking on average two flights per year). That roughly correlates to one return trip for each EU person a year. **In other words, the base rate for PNR data can be reasonably assumed to be in the region of 500 million.**

The Commission report and the staff working document appear to imply – and certainly do nothing to refute – that the 0.11% are all "true positives". However, that glaringly fails to take account of the base rate, and its impact on results.

To expand on the EDRi/Epicentre example, above: even if the PNR checks had a failure rate of just 0.1% (meaning that (1) in relation to persons who are actually terrorists or serious criminals, the PIUs (the national Passenger Information Units charged with "identifying" possible terrorists and serious criminals from the PNR data) will rightly confirm this as a proper "hit" 99.9% of the time, and fail to do so 0.1% of the time and (2) in relation to persons who are not terrorists, the PIUs will rightly not generate a confirmed "hit" 99.9% of the time, but wrongly register the person as a confirmed "hit" 0.1% of the time) the probability that a person flagged by this system is actually a terrorist would still be closer to 1% than to 99%.

In any case, even if the accuracy rate of the PNR checks were to be as high as this assumed 99.9% (which of course is unrealistic), that would still lead to some 500,000 false positives each year.

Yet the Commission documentation is silent about this.[12]

ii. Built-in biases[13]

It has long been known that algorithm-based assessments can be biased, even when they are entirely rational. As Marie Georges and I explained in our 2015 report, in which we also already referenced the (then still only proposed) PNR Directive:[14]

---

[12] In my opinion, I noted that the above shows that the base-rate fallacy is of absolutely fundamental importance to any discussion of the use of PNR data. I then returned to this in section 5 of that opinion, with reference also to the question of what should, and what should not, be regarded as a "true positive" in that context. That again is a good illustration of the general issues. I refer the reader of this excerpt to that section in the opinion.

[13] For more detailed discussion, see:

- Amnesty International, Netherlands: We sense trouble: Automated discrimination and mass surveillance in predictive policing in the Netherlands (AI Index Nr. EUR 35/2971/2020), 29 September 2020, available at: https://www.amnesty.org/en/documents/eur35/2971/2020/en/
- European Network Against Racism (ENAR), Data-Driven Policing: the hardwiring of discriminatory policing practices across Europe (authors: Patrick Williams And Eric Kind), November 2019, available at: https://www.enar-eu.org/IMG/pdf/data-driven-profiling-web-final.pdf
- Fair Trials, Automating Injustice: the use of artificial intelligence & automated decision-making systems in criminal justice in Europe, 2021, available at:
- https://www.fairtrials.org/sites/default/files/publication_pdf/Automating_Injustice.pdf

[14] Douwe Korff and Marie Georges, Passenger Name Records, data mining & data protection: the need for strong safeguards, (footnote **Error! Bookmark not defined.**, above), pp. 26 – 27.

Apart from the base rate fallacy (which is well-known to statisticians, albeit ignored by too many others …), the wider implications of algorithm-based decision-making have not been as widely researched as they should be. However, the leading research in this area, by Oscar Gandy, shows that (in David Barnard-Wills paraphrase):[15]

> predictive techniques and 'rational discrimination' – statistical techniques used to inform decision making by 'facilitating the identification, classification and comparative assessment of analytically generated groups in terms of their expected value or risk' – perpetuate and enforce social inequality.

This built-in risk - that profiles will perpetuate and reinforce societal inequality and discrimination against "out-groups", including racial, ethnic and religious minorities – is of course especially acute in relation to the screening of [airline passengers].

Crucially, this can happen even if the algorithms used are in their own terms perfectly "reasonable" and indeed rational. In practice (as Gandy has shown) the results will still reinforce the inequalities and discrimination already perfidiously embedded in our societies. **Crucially, this discrimination-by-computer does not rest on the use of overtly discriminatory criteria, such as race, ethnicity or gender (which is why the "anti-discrimination" clauses in the EU-US PNR Agreement, and indeed in the proposed EU PNR Directive, are so deficient, as discussed [later in the paper and below], under the heading "*Profiling and 'sensitive data'*")**. Rather, discrimination of members of racial, ethnic, national or religious minorities, or of women, creeps into the algorithms in much more insidious ways, generally unintentionally and even unbeknown to the programmers.[16]

But it is no less discriminatory for all that. Specifically, it is important to stress that in international human rights law, the concept of discrimination does not imply some deliberate discriminatory treatment. Rather, in the words of the Human Rights Committee established under the UN Covenant on Civil and Political Rights:[17]

> the term "discrimination" as used in the Covenant should be understood to imply **any distinction, exclusion, restriction or preference** which is based on

---

[15] Review of Gandy's main book on the topic, Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage, 2009, in *Surveillance & Society* 8(3): 379-381, at: http://www.surveillance-and-society.org/ojs/index.php/journal/article/viewDownloadInterstitial/gandy_chance/gandy_chance
For the book itself, see: http://www.ashgate.com/isbn/9780754679615 [original footnote]

[16] See in particular Fair Trials, Automating Injustice: the use of artificial intelligence & automated decision-making systems in criminal justice in Europe (footnote 13, above), section 2.1. For a particularly egregious example of bias in a widely used US system to predict re-offending, COMPAS, see: Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks, by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica, 23 May 2016, available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[17] UN International Covenant on Civil and Political Rights, Human Rights Committee, General Comment No. 18: Non-discrimination, 10 November 1989, para. 7, emphases added, available at: http://www.unhchr.ch/tbs/doc.nsf/%28Symbol%29/3888b0541f8501c9c12563ed004b8d0e?Opendocument
The HRCtee's definition draws directly on the definitions of discrimination against women, and discrimination on the basis of race, in the major UN Conventions against discrimination against women (CEDAW) and against people on the basis of race (CERD) (and, we might add, in the UN Declaration against discrimination on the basis of religion). [original footnote]

> any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status, and which has **the purpose or effect** of nullifying or impairing the recognition, enjoyment or exercise by all persons, on an equal footing, of all rights and freedoms.

As we add under that heading "*Profiling and 'sensitive data'*":[18]

> [T]he provisions in the EU-third country PNR agreements (in particular in the 2012 EU-US PNR Agreement) and in the … EU PNR Directive do not really seek to prevent discriminatory outcomes of the uses of the PNR data they claim to regulate. All they do is limit (to a rather limited degree) the overt use of such data.
>
> The problem is that, as already noted, datamining and profiling almost inevitably "perpetuate and [re-]enforce social inequality" – and that this is so, irrespective of any overt limitations on the use of "sensitive data": **you can use entirely "non-sensitive" data in such operations, yet still end up with results that in effect discriminate on grounds of race, religion or sexuality etc..**
>
> Given that stigmatisation of "suspect communities" is one of the most serious dangers of any state datamining/profiling operation, no more so than in relation to terrorism, the insufficiency of the safeguards in this respect in the PNR-related instruments is another major issue of concern.

This clearly applies to the PNR Directive. At first glance, it would appear that the directive seeks to ensure that the use of the "pre-determined criteria"/"profiles" will not result in discrimination. The relevant provisions are, in particular, Articles 6(4), 7(6) and 13(4), as elaborated on in recitals 15 and 20. They read as follows:

> **Article 6(4):**
>
> **Any assessment** of passengers prior to their scheduled arrival in or departure from the Member State carried out under point (b) of paragraph 3 **against pre-determined criteria shall be carried out in a non-discriminatory manner**. Those pre-determined criteria must be targeted, proportionate and specific. Member States shall ensure that those criteria are set and regularly reviewed by the PIU in cooperation with the competent authorities referred to in Article 7. **The criteria shall in no circumstances be based on a person's race or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership, health, sexual life or sexual orientation.**
>
> **Article 7(6):**
>
> The competent authorities shall not take **any decision that produces an adverse legal effect on a person or significantly affects a person** only by reason of the automated processing of PNR data. Such decisions **shall not be taken on the basis of a person's race or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership, health, sexual life or sexual orientation.**
>
> **Article 13(4):**
>
> **Member States shall prohibit the processing of PNR data revealing a person's race or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership,**

[18] Douwe Korff and Marie Georges, Passenger Name Records, data mining & data protection: the need for strong safeguards, (footnote **Error! Bookmark not defined.**, above), p. 37, original emphasis, cross-reference omitted.

> **health, sexual life or sexual orientation. In the event that PNR data revealing such information are received by the PIU, they shall be deleted immediately.**
>
> **Recital 15:**
>
> **[The list of the PNR data to be obtained by a PIU] should not be based on a person's race or ethnic origin, religion or belief, political or any other opinion, trade union membership, health, sexual life or sexual orientation.** ...
>
> **Recital 20:**
>
> [**No decision that produces an adverse legal effect on a person or significantly affects that person**] **should discriminate on any grounds such as a person's sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation**. The Commission should also take those principles into account when reviewing the application of this Directive.

A staff working document accompanying the European Commission report on the operation of the PNR Directive discusses these provisions as follows:[19]

> **Prohibition of processing of sensitive data**
>
> **The Directive prohibits the processing of 'sensitive data'** – that is, information which could reveal a person's race or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, health, sexual life or sexual orientation (Article 13.4). **In addition, the criteria against which PNR data can be processed, cannot be discriminatory and shall, in no circumstances, be based on a person's race or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership, health, sexual life or sexual orientation** (Article 6.4). **The principle of non-discrimination also applies to decisions made by national authorities, following the processing of PNR data** (Article 7.6).
>
> The prohibition of processing sensitive data was fully transposed by a large majority of Member States, with only two exceptions. In addition, all Member States but one require an immediate deletion of such data, if collected. However, **four Member States have failed to transpose correctly the prohibition on the use of discriminatory pre-determined criteria or criteria based on sensitive data. Five Member States did not transpose the obligation that decisions of competent authorities must respect the principle of non-discrimination.**
>
> With regard to the practical realisation of the prohibition to collect and process sensitive data, national authorities report that the IT systems of the Passenger Information Unit are designed in a way that makes the collection and processing of sensitive data technically impossible. This means that such data, even if transferred by air carriers, is filtered out and

---

[19] Respectively:
- Commission Staff Working Document accompanying the Commission Report mentioned below, SWD(2020) 128 final, 24 July 2020 (hereafter: "**Staff working document**"), section 4.7, available at:
- https://op.europa.eu/en/publication-detail/-/publication/9c419b94-cda3-11ea-adf7-01aa75ed71a1/language-en/format-PDF/source-search
- Report from the Commission to the European Parliament and the Council on the review of Directive 2016/681 on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime, COM(2020) 305 final, 24 July 2020 (hereafter: "**Commission report**"), available at: https://op.europa.eu/en/publication-detail/-/publication/4bfd0de3-cda3-11ea-adf7-01aa75ed71a1/language-en

> blocked or deleted by the system. In addition, **the fact that sensitive data are not collected in practice excludes the possibility of designing pre-determined criteria based on a person's race or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership, health, sexual life or sexual orientation.**
>
> (emphases added)

The latter claim is repeated later on in the staff working document:[20]

> The processing against pre-determined criteria is also limited by important safeguards. **The criteria** used must be targeted, proportionate and specific to the aim pursued and are subject to regular review. They **cannot be based on sensitive data and the assessment cannot be carried out in a discriminatory manner.** This limits the risk that discriminatory profiling will be carried out by the authorities.

It does not inspire confidence in the way the PNR Directive is applied in practice that four Member States have failed to transpose correctly the prohibition on the use of discriminatory pre-determined criteria or criteria based on sensitive data, and that five Member States did not transpose the obligation that decisions of competent authorities must respect the principle of non-discrimination.

But even if one leaves that aside, the document is **wrong** when it claims that "*the fact that sensitive data are not collected in practice [by the PIUs] excludes the possibility of designing pre-determined criteria based on a person's [sensitive data]*". This is for two reasons. First, there is nothing in the PNR Directive that bars the use of other data than the PNR data listed in Annex I to the directive in the creation of "pre-determined criteria" (in particular, if this is done in collaboration with other agencies). Indeed, it would appear that even the use of profiles created by other agencies (and that may be based on sensitive data, or proxies of such data) is not prohibited under the directive.

Second, the above claims are also **disingenuous**, in that they merely refer to avoiding the possibility of "*designing pre-determined criteria based on a person's [sensitive data]*" and of "*bas[ing] the criteria on sensitive data*". The suggestion is that if those criteria are not "based on" sensitive issues such as race, religion, etc., the application of those criteria will not result in discrimination, and indeed that no sensitive information will ever be revealed by the system:[21]

> [T]he PNR Directive strictly prohibits the processing of sensitive data. This constitutes an important difference with regard to the pre-existing draft EU-Canada PNR agreement, as explicitly acknowledged by the Court of Justice in Opinion 1/15. **The more intimate part of private life therefore remains fully protected by the processing operations provided for in the PNR Directive. Under the Directive, the information revealed by the processing of PNR data is in fact limited to the circumstances of the passenger's travel and would be established on the basis of data provided by the passengers themselves.**

That is a fundamentally misleading suggestion: as noted above, it has been shown time and again that biases can enter into algorithms and profiles even if they are not (directly) based on sensitive data – especially if proxies for such data are (knowingly or inadvertently) used.

---

[20] *Idem*, section 5.1, under the heading "*Additional safeguards surrounding the processing of PNR data*", on p. 18.

[21] *Idem*.

A typical example is the "red-lining" of districts: the labelling of individuals and households in certain areas (as defined, e.g., by postcode) as constituting a "high risk" in lending terms – which has resulted in discrimination of ethnic groups concentrated in such areas.[22] Highly sensitive matters can also be deduced from seemingly innocuous data:[23]

> Machine learning can ascertain a lot about you — including some of your most sensitive information. For instance, it can predict your sexual orientation, whether you're pregnant, whether you'll quit your job, and whether you're likely to die soon.

In simple terms: since "*intimate part[s] of [a person's] private life*" can be deduced, or at least inferred, from seemingly innocuous information – such as data included in PNRs (in particular if matched against other data) – those "*intimate aspects*" are <u>not</u> "*fully protected by the processing operations provided for in the PNR Directive*".

Indeed, in a way, the claim to the contrary is absurd: the whole point of "risk analysis" based on "pre-determined criteria" is to discover unknown, indeed hidden matters about the individuals who are being profiled: inferring from the data on those people, on the basis of the application of those criteria, that they are persons who "may be" involved in terrorism or other serious crimes surely is a deduction of an "intimate aspect" of those persons (even if it is not specifically or necessarily a sensitive datum in the GDPR sense – although if the inference was that a person "might be" an Islamist terrorist, that would be a [tentatively] sensitive datum in the strict sense).

Moreover, even without specifically using or revealing sensitive information, the <u>outcomes</u> of algorithmic analyses and processing, and the application of "abstract", algorithm/AI-based criteria to "real" people can still lead to discrimination.

The issue was recently addressed in another EDRi report, prepared at Delft Technical University, that stressed that there is:[24]

---

[22] See CBS, *Redlining's legacy: Maps are gone, but the problem hasn't disappeared*, 12 June 2020, available at: https://www.cbsnews.com/news/redlining-what-is-history-mike-bloomberg-comments/
As the article explains, although the practice was formally outlawed in the USA in the 1970s, "some housing advocates and lawyers say the practice continues, though in different form".

[23] Montreal Ai Ethics Institute, *When Algorithms Infer Pregnancy or Other Sensitive Information About People*, 2 November 2020, available at:
https://montrealethics.ai/when-algorithms-infer-pregnancy-or-other-sensitive-information-about-people/
The title is a reference to an (in)famous (although possibly apocryphal) anecdote about a father learning that his teenage daughter was pregnant due to the advertising company Target sending her coupons for baby items on the basis of inferences drawn from her shopping.

[24] EDRi, *Beyond Debiasing: Regulating AI and its inequalities*, report by Agathe Balayn and Seda Gürses, Delft University of Technology, the Netherlands, September 2021, p. 24, available at:
https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf
The report usefully clarifies some crucial terms:
"An **algorithm** is a process or set of rules to be followed to perform a calculation.
Machine learning algorithms are the set of calculations to perform in order to produce a machine learning *model* that will perform inferences regarding the future (e.g. predicting whether an individual is likely to recidivate).
These calculations are usually made on a set of training data: *essentially, the machine learning algorithm identifies the main patterns in available data and guides the learning of an inference behaviour that copies and amplifies these patterns.*" [*Continues overleaf*]

> **[a] conceptual difference between outputs and outcomes.** The outputs of the systems are the inferences they make on new data. Yet, the systems are always used in an environment where these outputs will impact things or stakeholders belonging to this environment.
>
> Thus, an outcome in this case refers to an output of a system and how it relates positively or negatively to a stakeholder.
>
> Bias and debiasing frameworks always consider the outputs of the systems, however we believe (and we will show this in the rest of the report) that considering the outcomes is more relevant when accounting for potential discrimination caused by the systems.
>
> (emphasis and underlining added)

Article 6(4) stipulates that **the assessment[s] of passengers prior to their scheduled arrival in or departure from the Member State** carried out under point (b) of paragraph 3 of the PNR Directive with the aim of identifying persons who require further examination by the competent authorities of the directive "***shall be carried out in a non-discriminatory manner***".

In my opinion, this falls considerably short of stipulating: (i) that the "pre-determined criteria" (the outputs of the algorithms) are not biased in some way and (ii) that measures must be taken to ensure that the outcomes of the assessments are not discriminatory. It is important to address both those issues, as the EDRi report just quoted stresses.

The issue is linked to the question of whether the processing is "high risk" and therefore requires an in-depth Data Protection Impact Assessment (or more broad human rights impact assessment). As Marie Georges and I put it in our paper, again with reference to Oscar Gandy:

> Only by constantly evaluating the results of the decisions based on profiles can one avoid these [discriminatory] effects. It takes serious effort. As Gandy concludes:[25]
>
> > these systems must be subject to active and continuous assessment and regulation because of the ways in which they are likely to contribute to economic and social inequality. This regulatory constraint must involve limitations on the collection and use of information about individuals and groups.
>
> In Europe, this "regulatory constraint" - this protection against discrimination-by-computer - takes the form of data protection rules (although, regrettably, to date not much action has been taken on this score).

A **machine learning model** refers to the output of this process of algorithm execution. Concretely, it is a set of mathematical equations with parameters learned from the data using the algorithm, and which can now be used to make inferences on new data following the patterns learned from the training data. …
When talking about the entities coming out of a machine learning model when presented with a data sample, the machine learning community often interchangeably uses the terms **prediction**, **inference**, **outcome**, and **output**." (p. 23, emphases added. Note the distinction made in the text between the last two terms.)

[25] Oscar Gandy, Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems, J Ethics Inf Technol, Vol 12, no. 1, pp. 29-42, 2010, at: http://academic.research.microsoft.com/Publication/41860489/engaging-rational-discrimination-exploringreasons-for-placing-regulatory-constraints-on-decision [original footnote]

The need for serious pre-evaluation of algorithms to be used in data mining and for continuous re-evaluation throughout their use is also stressed in various paragraphs in a 2021 Council of Europe recommendation on profiling,[26] e.g:

> 2.5 Member States should encourage the design and implementation of procedures and systems in accordance with privacy and data protection, already at their planning stage (privacy by design) and for the whole duration of data processing, notably through the use of privacy-enhancing technologies. ,,,
>
> 2.6 Profiling must not result in discrimination against individuals, groups or communities. It must undermine neither the dignity of persons nor democracy.
>
> …
>
> 2.10 The use of automated decision-making systems based on AI technologies poses additional risks due to possible errors and biases, and the difficulty of making the justification for decisions taken and ensuring transparency, consequently impeding the full exercise of the rights of the data subjects. The design, development and implementation of automated decision-making systems based on AI require special and continuous attention with regard to the risks created, and their assessment by multidisciplinary, independent teams.
>
> …
>
> 3.10 Appropriate measures should be taken by the controllers and, where applicable, the processors to correct data inaccuracy factors and limit the risks of errors and biases inherent in profiling.
>
> 3.11 The controllers and where applicable, the processors should periodically and within a reasonable time re-evaluate the quality of the data and of the statistical inferences used, as well as the impact of the use of profiling on the data subject's rights.

Here, I must also note that no serious efforts have been made by the European Commission or the EU Member States to fulfil these duties. Neither have ensured that full, appropriate basic information required for such serious *ex ante* and *ex post* evaluations is even sought or recorded.[27]

In sum: the European Commission and the EU Member States have not ensured that in practice the processing of the PNR data, and the linking of those data to other data (databases and lists), does not have discriminatory outcomes. The mere stipulation that outputs of algorithmic/AI-based profiling should not be "solely based on" sensitive aspects of the data subjects (the airline passengers) falls far short of ensuring compliance with the prohibition of discrimination.

---

[26] Council of Europe, Recommendation CM/Rec(2021)8 of the Committee of Ministers to member States on the protection of individuals with regard to automatic processing of personal data in the context of profiling, adopted by the Committee of Ministers on 3 November 2021 at the 1416th meeting of the Ministers' Deputies available at: https://search.coe.int/cm/pages/result_details.aspx?objectid=0900001680a46147

[27] I discuss this stubborn refusal to provide, or even look for, evidence of the effectiveness (or otherwise) of profiling and data mining in section 5 o my PNR opinion, that I am also reproducing as a separate blog.

iii. Opaqueness and unchallengeability of algorithm-based decisions including algorithm-generated "hits" (even if "confirmed" in a manual check)

In our 2015 report, Marie Georges and I also discussed the third issue with algorithmic/AI-based profiling:[28]

> [I]n the more developed "artificial intelligence" or "expert" systems, the computers operating the relevant programmes create feedback loops that continuously improve the underlying algorithms – with almost no-one in the end being able to explain the results: the analyses are based on underlying code that cannot be properly understood by many who rely on them, or even expressed in plain language.
>
> …
>
> increasingly, a state agency [may] place[ ] you on a terrorist "no-fly" or "high-risk" list, "because the computer said so": because the computer generated a "score" based on a profile, that exceeded or did not reach some predetermined basic level. If you ask for an explanation (if, that is, you actually find out that such an automated decision has been made on you), the … agency (or at least the person you are dealing with) is likely to be unable to explain the decision in any meaningful way. They might provide you with examples of some of the information used (e.g., that you used a "suspicious" route, or booked your ticket from a "suspicious" travel agent), but they will not give you the underlying algorithm - partly because the official him- or herself does not know or understand that algorithm, which is in any case constantly dynamically changing, and partly because the algorithm is a "national or commercial secret".
>
> **It is extremely difficult to provide for serious accountability in relation to, and redress against, algorithm-based decisions generally.**

In our report, we referred to the fact that in relation to certain "high-risk, rules-based" (i.e., algorithmic/AI-based) terrorists lists used by the US authorities and in particular in relation to their "no-fly" lists, the relevant supervisory agency, the US Government Accountability Office (GAO), had reported that even:[29]

> the procedures [aimed at] mitigat[ing] impacts [of these lists] on passengers who may have been misidentified to these lists [i.e., who had been wrongly labelled "high risk" and barred from boarding flights] [are] considered sensitive security information –

and those "procedures" were therefore not disclosed to the barred travellers. As we already concluded then: "*That leaves those who have been thus 'misidentified' (i.e., wrongly placed on 'no-fly' lists) without redress*". We added that:[30]

> Even at a higher accountability level, e.g., in relation to parliamentary or judicial or special oversight bodies, it will be effectively impossible to verify the risks inherent in those profiles:

---

[28] See Douwe Korff and Marie Georges, Passenger Name Records, data mining & data protection: the need for strong safeguards, (footnote **Error! Bookmark not defined.**, above), pp. 28 – 33. The quotes are from pp. 28 and 30 – 31 (with minor edits including references to "no-fly" lists added: the full paper explains that that is what the "misidentifications" relate to).

[29] *Idem*, p. 32, with reference to section I.ii of that report, in which the "high-risk, rule-based" lists and the GAO Report are discussed in further detail.

[30] *Idem*.

> i.e., to assess the level of "false positives" and "false negatives", or the possibly discriminatory effect of the profiles on certain groups, without the full, in-depth cooperation of the agency generating the profiles. Yet the latter are likely to be unwilling to be so helpful, unless compelled to do so by law.
>
> Profiling thus really poses a serious threat of a Kafkaesque world in which powerful agencies (like the DHS and the NSA – or in the near future European agencies?) take decisions that significantly affect individuals, without those decision-makers being able or willing to explain the underlying reasoning for those decisions, and in which those subjects are denied any effective individual or collective remedies.
>
> **That is how serious the issue of profiling is: it poses a fundamental threat to the most basic principles of the Rule of Law and the relationship between the powerful and the people in a democratic society.**

The issue is well illustrated by this very recent report from the UK:[31]

> **Disabled people are being subjected to stressful checks and months of frustrating bureaucracy after being identified as potential benefit fraudsters by an algorithm the government is refusing to disclose**, according to a new legal challenge.
>
> A group in Manchester has launched the action after mounting testimony from disabled people in the area that they were being disproportionately targeted for benefit fraud investigations. Some said they were living in "fear of the brown envelope" showing their case was being investigated. Others said they had received a phone call, without explanation as to why they had been flagged.
>
> **The Department for Work and Pensions (DWP) has previously conceded that it uses "cutting-edge artificial intelligence" to track possible fraud but has so far rebuffed attempts to explain how the algorithm behind the system was compiled.** Campaigners say that once flagged, those being examined can face an invasive and humiliating investigation lasting up to a year.

Similar stories relating to unfair, biased but opaque algorithms denying people – and especially people of colour – welfare benefits have emerged in the Netherlands and Denmark.[32]

These cases are of particular relevance because, as in relation to PNR data, these systems of course "merely" seeks to "identify" people who "may be" involved in crime (in that case, in fraudulently claiming state benefits), and this is then followed by "further investigation".

---

[31] *DWP urged to reveal algorithm that 'targets' disabled for benefit fraud*, Guardian, 21 November 2021, emphasis added, available at:
https://www.theguardian.com/society/2021/nov/21/dwp-urged-to-reveal-algorithm-that-targets-disabled-for-benefit?CMP=Share_iOSApp_Other

[32] See, e.g.:
Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms, Amnesty International, 25 October 2021, available at:
https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/
This Algorithm Could Ruin Your Life – A system used by the Dutch city of Rotterdam attempted to rank people based on their risk of fraud. The results were troubling, Wired, 6 March 2023, available at:
https://www.wired.com/story/welfare-algorithms-discrimination/
How Denmark's Welfare State Became a Surveillance Nightmare, Wired, 7 March 2023, available at ($):
https://www.wired.com/story/algorithms-welfare-state-politics/

Also noteworthy is the consistent refusal of those developing or using the systems to be open about them. The UK Department for Work and Pensions "*rebuffed attempts to explain how the algorithm behind the system was compiled*". In relation to the US re-offending algorithm, COMPAS, the company that developed the system, Northpointe, first refused to disclose any details, then provided some basic information and a set of 137 questions that were used to calculate an offender's likelihood of re-offending – which turned out to be highly dubious in terms of predictive value. Yet it still refused to disclose the way it weighed the answers and calculated the relevant scores – as always claiming that this was "proprietary [information]".[33]

In the context of the use of PNR data under the PNR Directive, the opaqueness of dynamic/self-learning algorithm/AI-based profiles has effects at several different levels:

- **PIU staff cannot challenge the computer output.** The opaqueness of reasoning underpinning the computer output makes it effectively impossible for the PIU staff to determine whether an initial "hit" against "pre-determined criteria" based on algorithm-based profiles is correct or not: all they can see is that the PNR data fed into the automated system generated a "hit"; since they do not know the underlying (ever-self-"improving") algorithm, there is no way they can check the validity or otherwise of this output.
- **The staff of the competent authorities are unlikely (or indeed also effectively unable) to challenge the computer output.** The opaqueness makes it difficult for the staff of any competent authority to which the (supposedly "confirmed" but not really checked: see above) "hit" is reported to validate or invalidate the "hit": they too really only know that the automated system generated a "hit", and it must be assumed that most of such staff also do not know and cannot really understand the underlying (ever-self-"improving") algorithm. In addition, there is the issue of what is known as "**confirmation bias**" that takes on a special form in relation to computer-generated outputs. The term "confirmation bias" refers to the general human tendency "to search for, interpret, favour, and recall information in a way that confirms or supports one's prior beliefs or values".[34] In relation to computer-generated suggested courses of action, humans have a tendency to trust the computer suggestion: they assume the data underpinning the computer reasoning (the algorithm) to be objective, fair and reliable – even if this will often not be the case: see above, at i and ii. Employees tasked with working with computer outputs have an additional reason to "go with" the computer output: if the computer output turns out the be wrong, they can blame the computer, but if they wrongly overrode the computer suggestion, they will be held responsible for the negative consequences.
- **Supervisory bodies cannot properly assess the systems.** External supervisory bodies such as Member States' data protection supervisory authorities will generally not be given access to the underlying data, cannot review the algorithms at the design stage or at regular intervals after deployment and in any case do not have the expertise. Internal

---

33 See Machine bias (footnote 16, above).

34 Wikipedia entry on *Confirmation bias*, available at:
https://en.wikipedia.org/wiki/Confirmation_bias
This refers to Raymond S. Nickerson, Confirmation Bias: A Ubiquitous Phenomenon in Many Guises, Review of General Psychology, 1998, Vol. 2, pp. 175 – 220, available at:
https://pages.ucsd.edu/~mckenzie/nickersonConfirmationBias.pdf

bodies are unlikely to be critical and may involve the very people who design the system (who write the code that provides the [dynamic] algorithm).

Article 7(1) and (2) of the Dutch PNR Law stipulates that:[35]

> [The pre-determined criteria] must be determined and regularly tested by the PIU in cooperation with the relevant competent authorities.
>
> [Those] criteria must be suitable for their purpose, proportionate and specific to the crime in relation to which, according to the criteria, the possible involvement of a person can be determined.

A report on the evaluation of the Dutch PNR Law notes that, to that end:[36]

> a special commission [has been established] that tests the pre-determined criteria (including the weighing [of the various elements] and the threshold value [for regarding a "hit" against those criteria to be a valid one]) against the requirements of Article 7(2) of the Law.

The commission consists of staff of the Dutch PIU and representatives of the relevant competent authorities and of the state prosecution office (*Openbaar Ministerie*, *OM*). It is unclear – and in my opinion doubtful – whether any of the members have special expertise in relation to the matters I have discussed: the base-rate fallacy, detecting bias in outputs *and outcomes*, or checking the evolution of self-learning algorithms. In any case, they are clearly not detached from the system, or independent or impartial.

The Data Protection Officer of the PIU is not a member of the commission, but "has access to the reports" of the commission. I doubt that those reports allow the DPO (or any expert they may consult, if that is allowed) to verify the suitability, effectiveness and proportionality of the criteria. Moreover, in any case, the report adds that:[37]

> **The rules [on the creation of the pre-determined criteria] do not require the weighing [of the elements] or the threshold value [for regarding a "hit" against those criteria to be a valid one] to meet objective scientific standards.**
>
> **This is quite an astonishing matter. It acknowledges that the algorithm/AI-based profiles are essentially unscientific. In my opinion, this fatally undermines the way the pre-determined criteria are created and "tested" in the Netherlands. Yet at the same time, the Dutch system, with this "special commission", is probably better than what is in place in most other EU Member States. This surely is a matter that should be taken into account in any assessment of the PNR system EU-wide.**

---

[35] Original Dutch text:

1. De criteria, bedoeld in artikel 6, eerste lid, onderdeel c, worden door de Passagiersinformatie-eenheid in overeenstemming met de betrokken bevoegde instanties vastgesteld en regelmatig getoetst.
2. De criteria zijn doelgericht, evenredig en specifiek voor het misdrijf waarbij de mogelijke betrokkenheid van een persoon overeenkomstig de criteria kan worden bepaald.

[36] Irion, K., Es, R. van, Meeren, K. van der, & Dijkman, D, Evaluatie PNR-Wet, WODC rapport 3181, October 2021, p. 105, available at:
http://hdl.handle.net/20.500.12832/3118
https://www.eerstekamer.nl/overig/20211112/evaluatie_pnr_wet_wodc_oktober/document

[37] *Idem*, emphasis added.

The only way to guard against erroneous or societally unacceptable outcomes of dynamic (self-learning) algorithm/AI-based processing and matching of data would be to have those algorithms and their application and the outcomes of their application *continuously rigorously tested and audited by fully qualified,* independent *experts on the basis of clear, peer-reviewed scientific standards in order to limit, as far as possible: straight-forward errors, bias against certain groups (especially those defined by race, gender, religion, etc.), and excessive false positives and/or false negatives. Moreover, in a democratic society the results of those tests and audits – and the underlying algorithms – should be* open to external, independent scientific review*, not least also on behalf of any individuals affected by the programs.*

At present, there is no conceivable way in which such checks and audits could be implemented: the political will is not there; the entities involved would forcefully oppose any such "interference" in their practices; and commercial entities involved in the creation of the algorithms would claim it would infringe their proprietary (intellectual property) rights. As I will show in section 5, sub-section 5.1, below, the agencies and the Member States' governments have deliberately withheld from even collecting the data that are necessary to judge the suitability, effectiveness and proportionality of their mass surveillance activities. They will not suddenly welcome openness, transparency and scientific testing of their new tools (even if, or perhaps especially because such transparency very likely would, show that those new tools are not suitable, not effective, and not proportionate to their legitimate aim).

In sum:

- **because the "base-rate" for the PNR data mining is so high (in the region of 500 million people) and the incidence of terrorists and serious criminals within this population so relatively low, algorithm/AI-based profiling is likely to result in tens and possibly hundreds of thousands of "false positives": individual air passengers who are wrongly labelled to a be person who "may be" involved in terrorism or other serious crime;**
- **the provisions in the PNR Directive that stipulate that no sensitive data may be processed, and that individual decisions and matches may not be "solely based on" sensitive aspects of the individuals concerned do not protect those individuals from discriminatory outcomes of the profiling;**
- **the algorithm/AI-based outcomes of the processing are almost impossible to challenge because those algorithms are constantly dynamically changed ("improved" through self-learning) and therefore in effect impossible to fully comprehend even by those carrying out the analyses/risk assessments; and**
- **the outputs and outcomes of the algorithm/AI-based profiling and data mining and matching are not subject to proper scientific testing or auditing, and extremely unlikely to made subject to such testing and auditing.**

And this applies essentially to all AI-based data mining to "identify" rare bad people.

- o – O -o -

Douwe Korff/Cambridge (UK), 2 May 2023