PHRP EXPERT MEETING ON PREDICTIVE POLICING

On 20.-21. May 2019, the Police and Human Right Programme (PHRP) of Amnesty International Netherlands held an expert meeting on predictive policing. The meeting brought together representatives of:

AlgorithmWatch (Germany); Amnesty International (The Netherlands and International Secretariat); Belgian Federal Police; Big Brother Watch (UK); Data Justice Lab (Cardiff University); Digital Freedom Fund (UK); Center for Security Studies (CSS), ETH Zurich; HUMAINT research project of the Joint Research Centre at the European Commission; European Union Agency for Fundamental Rights; Helsinki Foundation for Human Rights; Liberty (UK); TNO (Dutch Organisation for Applied Scientific Research); the Department of Criminal Law, University of Valencia; School of Computing, University of Utah; the West Midlands Police and Crime Commissioner.

Introduction

Technological developments increasingly find their ways into today's policing, and artificial intelligence (AI) is one of them. Police use data sets of different sizes to feed into an algorithmic model that is supposed to predict either places where crime is most likely to occur in the near future (place-oriented predictive policing), or persons who are likely to get involved in crime (person-oriented predictive policing). While such models are more and more developed and used by police agencies across the globe, there are many questions with regard to the accuracy of the outputs, possible discriminating biases in the underlying data sets and / or the model using the data, their effectiveness to actually predict crime and many other elements more. It is particular important to answer these questions given the great human rights impact the use of such technology can have with regard to data protection and the right to privacy, the right to liberty and security, freedom from discrimination, freedom of expression and information, the right to a fair trial and effective remedy etc.

The PHRP organized this meeting with a view to bring experts from different areas (police, criminology, data scientists, academic researchers, civil society) together in order to discuss some key questions in this area. This report summarizes the main elements of the discussion with a view to nurture the reflection about critical problems to be researched in depth and to be addressed – in particular from a human rights perspective.

Looking at predictive policing methods used across Europe[1] and the United States, the meeting revealed a wide range of different approaches: considering different types of crime (burglary, violent crime, illegal immigration, child abuse etc.); being person- or place-oriented; differing in the type and variety of data used (e.g. data exclusively linked to crimes committed; sensitive data such as information from the social welfare system, income, ethnicity, nationality or gender of individuals or

[1] AlgorithmWatch, *Automating Society: Taking Stock of* Automated Decision-Making *in the* EU, January 2019, available at: *https://algorithmwatch.org/wp.../01/Automating_Society_Report_2019.pdf*.

of the population resident in a given area; using only very few features to several thousands); the use or not of specific technical means such as facial recognition, automated number plate recognition, a self-learning algorithm, or even some first attempts of behavioural recognition etc.

Depending on the approach, questions regarding a clearly defined operational purpose, accuracy, control and transparency (in particular when private companies were involved in the design and/or the operation of the underlying algorithm) as well as the involved human rights impact would also get very varying answers. Comparing for instance predictive policing approaches and possible problems related to them in Europe and in the US seems to be very difficult, due to the largely differing legal systems and policing context and history. Nevertheless, a range of critical elements were found to require attention in any context:

PLACE-ORIENTED PREDICTIVE POLICING

Data quality

Participants shared a quite critical view on the quality of the data used in predictive policing. Overall, any model is only accurate if the input data to train the model is drawn from the same distribution as the world the model is applied on. However, data input used to train and feed into the algorithm is influenced by the data that is actually available: Input data can be outdated, incomplete or not representative. It can even be biased as a result of a biased policing approach in the past, with police focussing on particular areas or communities, or be otherwise inaccurate (e.g. due to over- or underreported crime, when mixing types of crimes that do not fit together, or in some cases even be based on deliberately fraught or falsified data).[2] Thus, participants agreed that the input data is never reflecting the full truth, but only a part of it. In particular, in many instances it will reflect the police's previous approach and priorities regarding certain areas, groups or specific types of crimes, approaches that may often be shaped by a structural bias.

The greatest concern discussed in this regard was when, as a result, the output is biased, and then leading to discriminatory over-policing of certain communities. A possible remedy proposed was to change the way the algorithm output is used: to move away from a decision-making tool (where and whom to police) to using it as a diagnostic tool: why is the output what it is? E.g. if a certain group of people or a certain area shows greater prevalence of crime, where does this come from? It may be an indicator of a biased policing approach focussing excessively on this group or area. This would allow police to critically reflect on their relationship with this community and how they could improve their approach. It could also help to look deeper into the causes of crime (e.g. poverty, lack of opportunities, social exclusion) and address them rather than to choose a law enforcement approach.

Quality of the algorithmic model

Other problems discussed related to the quality of the model: There are various types of data that can be included in the design of an algorithm. The challenge, however, is how to identify those features that are statistically relevant for determining anticipated crime and how much weight should be given to each feature. These decisions are not taken by crime specialists, e.g. criminologists or police officers, it is usually taken simply based on purely statistical analysis: A large

[2] Richardson, Rashida and Schultz, Jason and Crawford, Kate, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice (February 13, 2019). New York University Law Review Online, Forthcoming. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423 .

variety of data is entered into a system that then establishes a statistical correlation. The information about this statistical correlation will then be used to guide policing decisions (e.g. when and where to patrol). Thus, in some cases there may be statistical correlations that do not have a causal link and lead to correlations irrelevant for the prediction of crime. The more features are entered into a system, the greater will be the inaccuracy of the system and the more difficult will it be to assess which features are actually relevant.

Besides statistical relevance, there is also an element of value judgement: even if some features might show a statistical correlation to crime, it is a different question whether it is ethically correct to include them in the algorithm, e.g. sensitive features such as ethnicity, religion, income. The CAS system in The Netherlands, seeking to predict burglaries, for instance, includes information such as average income or reception of social welfare benefits in the input data. The category of "Non-western immigrants" was initially included but deleted in 2017. In that regard, it was stressed that the designing of an algorithm needs to be guided by a clear purpose and transparent values of what is supposed to be achieved with it. A problem mentioned in this regard was also the fact that data might be gathered for a specific purpose and will then later be used for different purposes – raising concerns in relation to data protection and privacy, but possibly also discrimination, since certain data will be mainly available from specific groups, e.g. immigrants or low income families.

Self-learning systems

Self-learning systems were found to present additional problems:

1) As soon as the systems receives updated information, the output does not reflect anymore the real world, but how the system sees the world, and this can be different. Further, since the system will weigh the information it gets, small differences at the input stage will be exacerbated throughout the process. Areas considered to have a slightly higher risk of crime than others will be highlighted over time as high-risk areas compared to others ranked low risk, while the difference in the beginning was much more nuanced. This means also that pre-existing biases will be exacerbated, contributing to even greater exposure, possibly even criminalising an entire group of people who are already victims of a historical biased policing approach.

2) The risk of feedback loops: when as a result police increase patrolling in one area, they are more likely to encounter criminal activity, which will feed back into the system, leading to even more policing in the future. One solution presented was reinforced learning of the algorithm to avoid such feedback loops: i.e. to include in the algorithm model as a corrective element the information about increased policing in a given area. Though, it seems that this option has not yet been implemented in practice and is only partially a solution: If a model is trained on one distribution, it will in general not work anymore if the underlying distribution changes (e.g. as a result of a change in the policing approach or of a change in resident behavior as a result of police presence). The monitoring of changes in the underlying distribution is very difficult and it was stated, that currently, the means are not available to ensure sufficient accuracy of these models.

Furthermore, it was also pointed out that systems and datasets will never be entirely bias free. It was repeatedly underscored that police officers need to be trained on how these systems work, including getting to understand the weaknesses, possible biases and risk of inaccuracy of the outputs. By acknowledging this and by making the system transparent, one can use predictive systems to identify bias both in the system itself as well as more generally in the approaches taken by police agencies.

Effectiveness of place-oriented predictive policing

Participants found it also almost impossible to evaluate whether place-oriented predictive policing actually helps to reduce crime and can generally be viewed as effective. Challenges were identified at different levels:

1) Already to ascertain that there is or not a reduction in crime can be difficult. There is the possibility of a "waterbed"-effect, where crime is actually only displaced, but not reduced. Then there can also be a normal fluctuation in crime rates.

2) Even if crime figures went down, it appears almost impossible to determine whether this the result of the measures taken or whether other factors lead to this result. Some studies on the PRECOBS system used in Germany have illustrated this difficulty: in some areas where the system was used, crime figures decreased, while in other areas this was not observed. At the same time, there was also a reduction in crime in areas where PRECOBS was not used. So, it could well be that crimes would have gone down even without PRECOBS. Further, predictive policing is often not used in isolation but in parallel to other policing measures. This makes it difficult to attribute decreasing crime rates to one specific measure or to determine predictive policing as a contributing factor.

3) Similarly, it is difficult to test to what extent predictions are accurate. If police increase patrolling in response to a prediction, and this increased patrolling has the desired effect: deterrence of crime, then – in the absence of crime occurring – one will never know if the prediction was actually correct. Actually, a possible way to assess the accuracy of the prediction would be to not act on predictions and see if they manifest, but this is seldom a feasible option in practice. However, it was also mentioned that police sometimes seem rather to be satisfied if there is a reduction in crime, even without the proof that it is a result of predictive policing.

4) Another question in that regard is how to define what is considered a success of predictive policing. Police may consider it successful if they carry out an arrest, even if it does not lead to a conviction. Further, crime reduction alone should not be considered a success, as it needs to be balanced against the costs for the community, e.g. regarding discrimination, over-/under-policing, relationship between police and the community, economic impact (e.g. the value of property, insurance rates) etc.

In view of these challenges, participants considered it indispensable, prior to decide on the use of predictive policing systems, to clearly define the goals of what is supposed to be achieved, and to thoroughly assess whether predictive policing is actually needed to achieve it. In some cases, predictive policing systems may not be necessary to achieve the defined goal. Simply looking at available data may already give enough insights to act on, without the need to feed the data into an algorithm and let a machine determine the way policing is done. Another underlying question should be one of police priorities, which includes a reflection on whether those crimes that can be predicted by systems are the crimes that police should focus on, and whether the negative impact on the community outweighs the benefits of deploying predictive policing. The use of data to enforce immigration policies in the UK was cited as a worrying example on how seriously this can impact on the daily life of an entire community, leading people to refrain from accessing health or other basic services.[3]

[3] Liberty, "Care, don't share", London, December 2018, available at: https://www.libertyhumanrights.org.uk/news/press-releases-and-statements/liberty-launches- %E2%80%9Ccare-don%E2%80%99t-share%E2%80%9D-campaign-and-report.

Participants expressed concern that a clearly defined goal is indeed often absent when building and deploying the algorithm. While predictive systems may be deployed by police with legitimate intentions to counter crime more effectively, there seems to be a perceived need to use new technologies without carefully evaluating whether they are actually beneficial. Also, while often argued for in the context of scarcity of resources and austerity measures, it was pointed out that predictive policing does not actually reduce the need for human resources. On the contrary, it increases the need for resources as it requires personnel to maintain and train the system, to assess the accuracy of outputs, to evaluate them in view of the possible operational measures to be taken, as well as officers to act on the predictions.

PERSON-ORIENTED PREDICTIVE POLICING

Data quality, data protection and privacy

Person-oriented policing systems can be based on a wide range of data, from police records to postcode data and health records. Databases may also include persons who have never been in contact with the criminal justice system themselves but appear in police reports because they were present when e.g. someone else has been arrested. Here again, issues of data protection and privacy are of great concern. Furthermore, a clearly defined purpose and strict regulations of when and how outputs can be used by whom are required. The Gang matrix of the Metropolitan Police in the UK is a very problematic example in this regard where data of persons, who had never been involved in any crime at all, were shared with a range of public institutions in the educational, health, employment and other sectors, which heavily affected the lives of these persons.[4]

Such systems often apply generalised and biased assumptions to individuals, e.g. because of being part of a certain minority group. Such bias cannot be avoided by simply excluding certain grounds, such as ethnicity, from the algorithm, as there are correlations with other features that are included in the algorithm (e.g. post codes of areas mainly inhabited by a migrant population). If there is a statistical correlation in the data between certain factors, such as nationality, and criminal behaviour, the algorithm will pick it up even if the feature is not explicitly considered by the algorithm. Again, the Gang Matrix mentioned above reveals such a problem: In October 2017, more than three-quarters (78 per cent) of those listed in the Matrix as being part of a gang were black, while according to police figures only 27 per cent of those responsible for serious youth violence are black. An evaluation of the COMPAS system used in the USA to determine the risk of recidivism of offenders also indicated racial bias, since black people were more likely to be considered high risk.[5]

Quality of the algorithmic model and bias detection

Another question discussed was how to evaluate the quality of an algorithm and to detect possible biases in the designed model. For instance, when looking at algorithms seeking to predict the risk how likely a person is to reoffend, the following discrepancies can be indicators of bias:

1) Unequal risk thresholds: As a result of the specific characteristic of a certain group of people and their ranking within this group, individuals in that group can be classified “high risk” at a score

[4] Amnesty International, United Kingdom, Trapped in the matrix, November 2018, available at: https://www.amnesty.org.uk/london-trident-gangs-matrix-metropolitan-police.

[5] ProPublica, Machine Bias, May 2016, available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

that is lower than for individuals in another group who score below average in their group, but in absolute numbers turn out to present a much higher risk.

2) The percentage of persons classified "high risk" of reoffending is higher in one group than in another group.

3) The percentage of false positives is higher in one group than in another group.

It is difficult to establish, however, whether such discrepancies reflect the reality or are evidence of a bias in the design of the algorithm. Equalizing input data (e.g. giving the same reoffending rate to all groups) might be a possible test to see if the algorithm would show the same results or differ; if the output remains the same, there might be underlying features within the model that lead to bias in the output.

In practice, however, the statistical risk established through the input data will usually differ across groups. Thus, it would have to be assessed whether this reflects the actual situation (difference in input data justified) or is just the result of policing bias in the past (difference in input data not justified). Simply equalising input data in practice (and not just for testing) in order to counter the differing output of the algorithm, however, is not possible since this might then turn out to be discriminatory towards the other group.

Accuracy of predictions

In terms of accuracy, predictive systems have the advantage that compared to human beings, the output is not influenced by cognitive bias and mood. Furthermore, using the example of psychiatric assessments of potential violent behaviour, it was illustrated that humans tend to overestimate dangerousness and their predictions can result in a high rate of false positives, i.e. people being wrongly categorized "high risk". Here, at a first glance, computing data seems to promise much greater accuracy. However, in many instances, evaluations of these systems tend to focus on the accuracy rate of the true positives (how many of those identified high risk actually committed a violent act) and not to look at the false positives (those identified high risk, but not having committed any act of violence) and for the latter the accuracy is often much lower than for the true positives. Furthermore, there has been a shift in focus to other statistical measures, such as relative risk, establishing rather whether a person has a greater or lower risk compared to another group. This, however, does not tell anything about how likely the person is indeed to commit a violent crime. It only compares the person to another group. As a result, these predictions – similar as human beings - turned out to rather give higher risk scores to a person, leading to a high rate of false positives. Still, all too often and looking only at true positives, computing systems are considered to be highly accurate.

Decision making

Another problem discussed was the fact that, the outcome of computed risk assessments is less accessible to people as, unlike assessments done by humans, comprehending them requires technical and statistical expertise. In the absence of fully understanding how the algorithm works, it is likely that the output of the system has a high impact on decision making. Furthermore, participants shared the concern that even if the computed risk assessment is accompanied by an explanation of how it came to the result (incl. the risks of inaccuracy and bias), a high-risk score is still likely to impact decision making. Judges, for instance, were said to be reluctant to release a person classified as "high risk" by an algorithm. Overriding the result of the system would put the person into a position where they would not only have to justify their decision but would likely be blamed if their decision was wrong, both for taking a wrong decision and for not listening to the algorithm.

There is thus a tendency to adhere to the predictions of systems, relieving oneself from assuming the responsibility for the decision made. However, there was consensus among participants that, if at all, predictive systems should only be one element feeding into human decision making, but not replace it. Furthermore, it was pointed out that empathy is an important factor in human decision making which is absent in an algorithm.

As a matter of principle, decisions taken in the context of person-oriented predictive policing must respect the principle of proportionality and here, the risks of false positives are particularly relevant. While it may never be possible to avoid false positives completely, the extent to which they can be considered acceptable should be proportionate to the consequences for the affected individual. If the actions taken by police as a result of a "positive" do not affect the rights of an individual, acting on a false positive might be acceptable compared to the costs of a false negative (i.e. failing to identify and act upon a possible risk). However, if the consequence would be a restriction of fundamental rights, such as in the case of a decision for or against imprisonment of a person, false positives are to be avoided at all costs: In such cases, it was suggested that in accordance with principles of criminal law, the concept of "*guilt to be proven beyond reasonable doubt"* should be adhered to. The purpose of criminal law is not to fight crime at any costs, and these fundamental principles should not be thrown overboard simply because technology is involved.

An underlying concern that was voiced in that regard was that such systems are based on an oversimplification of crime and the factors that lead to criminal behaviour. Thus, it was suggested, that instead of serving as a decision making tool, risk-assessment tools can be useful to analyse which factors contribute to criminality, in order to address the underlying causes not at the level of the individual but in a structured way, with a focus on prevention programs.

ACCOUNTABILITY

Individuals as well as the community as a whole have the right to know, understand and challenge decisions that affect their lives and human rights. This includes the right to effective remedy, whenever the state takes decisions that infringe on human rights of a person. However, participants agreed that this proves to be particular difficult when it comes to the use of algorithms in decision making. In many situations, people may not know at all that an algorithmic model was applied and has an impact on their lives. And even if they learn about this, it will be difficult for them to challenge any decision taken, given the obscurity around the use of this technology.

Throughout the meeting, there was consensus that authorities must ensure transparency and accountability when deciding to resort to the use of algorithms in policing. This would first of all require a solid legal framework regulating the use of predictive policing systems including a clear definition of the operational goals, the proceedings for developing the methodology (incl. type of data input, proper evaluation and quality control, prevention of bias and discrimination, layered access rights etc.), as well as the decisions that may or not be taken in view of the algorithm output. Already existing legal norms (e.g. regarding the presumption of innocence, privacy and data protection) must also be applicable and adequately applied to these systems. In addition, participants discussed the option of introducing an ethical framework in order to ensure a greater human rights compliance in the development and use of such systems. When deciding to use such systems, law enforcement agencies should be subject to clear ethical rules to be drafted in close consultation with civil society.

Especially when the underlying computing systems are developed by private companies, there is however a risk that subscribing to ethical standards turns into a mere window dressing exercise due to a lack of incentive, and it would need to be ensured that companies can be held accountable to the standards they subscribe to. In addition, software developed by private companies is often not transparent due to trade secrets. Public institutions in turn might evoke security or other concerns when asked to ensure transparency of the algorithm. It was suggested, that access to the algorithms might not be granted not to the public as such but to specific oversight bodies, which would allow to address such concerns and allow companies to maintain their competitive advantage. Another option to address this problem might be resorting to non-profit organisations developing the algorithm, an option that is now sometimes chosen in the USA, and that might provide for greater transparency. What should in any way be public is which systems are in use by the authorities, for what purpose they are being deployed, what data they run on and what costs are involved.

Regarding transparency and accountability, participants highlighted the particular challenge of self-learning systems: they are very difficult to be assessed since at a certain stage, even the designer of the model might not know anymore how the systems works as a result of the system's self-modification through the feedback input received. Other methods of testing such self-learning systems regarding their accuracy and absence of bias are therefore required, but not yet in place.

Participants agreed that for individuals affected by algorithmic decisions, an effective right to remedy can only be exercised if one has the necessary information to challenge a suspected violation. This includes transparency in who is being affected by a system, e.g. the knowledge that one's name is part of a police listing. They should further get an explanation of how a certain decision was made, in particular if it is done solely on the basis of an automated output, or if to which extent there was involvement of a human being.

More difficult to answer is then the question what to challenge precisely: The computed score, which is placed on an individual may be – mathematically – correct. Thus, the question is, whether it will be possible to challenge the use of the algorithm as such, including the fact that a person is assessed not according to his or her individual situation, but according to an average score attributed to him or her, or to challenge the decision taken on the basis of the risk score. The latter can be particularly difficult since, even if there is a human being such as a judge who pronounces a decision, it may still be the average score that will be the determining element for that decision. In any case, those seeking to claim that their rights have been violated would need to be given all the relevant information about a decision taken, and when this decision is taken on the basis of an output from an automated system, they should be entitled to an explanation of how the system got to the specific output, in order to challenge it.

Finally, for this challenging of automated outputs and decision making to be effective, participants considered it particularly important for judges to familiarize themselves with new technologies and become digitally literate in order to be able to evaluate whether the human rights of a person have been violated through the use of such a system or not.

CONCLUSION AND OUTLOOK

As demonstrated through this report, the meeting revealed a wide range of areas of concern and in need for further research and challenging debate. For PHRP, the following areas will be given particular attention in its upcoming research:

- To which extent there is a remedy to the problems of data quality
- To which extent there is a technical remedy to problems in the functioning of the algorithm (reinforced learning, testing accuracy and possible bias in algorithm models through variations in data input etc.)
- To challenge more systematically the effectiveness of predictive policing approaches to actually prevent crime (in particular in relation to hotspot policing).
- The general acceptability of person-oriented predictive policing from a human rights perspective, in particular in view of the human rights impact of false positives
- To identify more clearly short comings in the design and implementation of predictive policing systems as well as the related human rights costs with a view to demonstrate the need for thorough processes rather than try-and-error-approaches.
- Looking at different ways of using algorithm models in policing (e.g. as a diagnostic tool instead of a decision-making tool)
- How to create effective accountability and transparency with a view to uphold human rights of persons affected by predictive policing approaches.
-

Anybody whishing to exchange with PHRP on these questions is warmly welcomed to contact us under: phrp@amnesty.nl.