Algorithmic risk assessment policing models: Lessons from the Durham Constabulary HART model

OSWALD, Marion, GRACE, Jamie <http://orcid.org/0000-0002-8862-0014>, URWIN, Sheena and BARNES, Geoffrey

Published version

OSWALD, Marion, GRACE, Jamie, URWIN, Sheena and BARNES, Geoffrey (2018). Algorithmic risk assessment policing models: Lessons from the Durham Constabulary HART model. Information and Communications Technology Law, 27 (2), 223-250.

M. Oswald (Winchester), J. Grace (SHU), S. Urwin (Durham Constabulary) & G.C. Barnes (Cambridge) (2017)

Executive summary

Algorithmic risk assessment policing models: Lessons from the Durham Constabulary HART model

Key findings

To permit the use of unproven algorithms in the police service in a controlled and time-limited way, and as part of a combination of approaches to combat algorithmic opacity, our research proposes 'ALGO-CARE', a guidance framework of some of the key legal and practical concerns that should be considered in relation to the use of algorithmic risk assessment tools by the police. As is common across the public sector, the UK police service is under pressure to do more with less, and to target resources more efficiently and take steps to identify threats proactively; for example under risk-assessment schemes such as 'Clare's Law' and 'Sarah's Law'. Algorithmic tools promise to improve a police force's decision-making and prediction abilities by making better use of data (including intelligence), both from inside and outside the force. This research uses Durham Constabulary's Harm Assessment Risk Tool (HART) as a case-study. HART is one of the first algorithmic models to be deployed by a UK police force in an operational capacity. Our research comments upon the potential benefits of such tools, explains the concept and method of HART and considers the results of the first validation of the model's use and accuracy. The research concludes that for the use of algorithmic tools in a policing context to result in a 'better' outcome, that is to say, a more efficient use of police resources in a landscape of more consistent, evidence-based decision-making, then an 'experimental' proportionality approach should be developed to ensure that new solutions from 'big data' can be found for criminal justice problems traditionally arising from clouded, non-augmented decision-making. Finally, our research notes that there is a sub-set of decisions around which there is too great an impact upon society and upon the welfare of individuals for them to be influenced by an emerging technology; to an extent, in fact, that they should be removed from the influence of algorithmic decision-making altogether.

Background: Durham Constabulary and the HART model

The 'Harm Assessment Risk Tool' (or 'HART') was developed by statistical experts based at the University of Cambridge in collaboration with Durham Constabulary. It has been developed to aid decision-making by custody officers when assessing the risk of future offending and to enable those suspects forecast as moderate risk to be eligible for the Constabulary's Checkpoint programme. Checkpoint is an intervention currently being tested in the Constabulary and is an 'out of court disposal' (a way of dealing with an offence not requiring prosecution in court) aimed at reducing future offending. For schemes where difficult risk-based judgements are required, it has been argued that a fair and trustworthy algorithmic decision-making tool may potentially be helpful, provided not used in a determinative way. All algorithmic responses use the past, where the outcomes have already taken place, as a model of what will take place in the future. The HART model is built using approximately 104,000 custody events over a five year period (2008-2012). It uses 34 different predictors to arrive at a forecast, most of which focus upon the prior offender's history of criminal behaviour.

Context and Findings

The data used as predictors in HART will, for the time being, remain limited to those held within Durham Constabulary systems. The system will not utilise data from other local agencies in Durham, other police force areas, or national IT systems such as the Police National Computer or the Police National Database. This limitation is just one reason that such models can serve only to *inform* human decision making, and will remain unable to function as the ultimate decision maker at any stage of the criminal justice system. The model simply does not have all of the information available to it, and can therefore only *support* human decision-makers, rather than replace them. The custody officers will long retain their discretion and the model will not fetter the options available to them. With both their own local knowledge and their access to other data systems, custody officers will frequently be aware of other information that overrides the model's predictions, and they must apply their own judgement in deciding upon the disposition of each offender's case.

An independent validation study was conducted of HART during 2016, with data not used to build the model. Custody data for the full year of 2013 were used for the validation, using just under 15,000 custody events. The model's forecasts for each custody event during 2013 were then compared to the actual, known outcomes over the following 24 months. The 2013 validated accuracy overall of the model was 62.8%, which reflects a modest drop from construction estimate of 68.5%. The largest loss of accuracy in validation occurred amongst those that had actual high risk outcomes, where the accuracy rates fell from 72.6% to 52.7%. In Durham Constabulary, the initial version of HART has required the custody officers to make their own predictions of each offender's future arrests whenever the algorithm has been used. These data will eventually allow a direct comparison of the police officer's human judgement to the HART forecasts. Early results show that custody officers are generally uneasy with forecasting at either extreme, and avoid making both high and low risk predictions. A substantial majority of officer predictions are for moderate risk behaviour (61.4%), and the model and officers agree only 55.5% of the time. There is a clear difference of opinion between human and algorithmic forecasts. Nevertheless, caution should be taken to not hold algorithms to an idealistic standard of accuracy that does not exist in reality.

The HART model contains over 4.2 million decision points, all of which are highly interdependent on the ones that precede them within the tree structure. These details could be made freely available to the public, but would require a huge amount of time and effort to fully understand. It is becoming increasingly difficult to explain to non-computer scientists and non-statisticians how a machine learning forecasting model arrives at its outcomes, and the potential for misunderstanding and even intentional misrepresentation is vast. Our argument is that a model must be developed (*as provided overleaf, below*) which ensure accountability and best-practice from the off.

Implicit in the points made in the 'Lawful' section of 'Algo-Care' below is whether a statistical, algorithmic method is appropriate *at all* in each given situation, and whether it can ever be justified to use certain categories of data, for instance ethic origin, as 'inputs'. We would advocate that, as part of a programme of legal regulation or police adoption of algorithmic intelligence analysis models, clarity is needed as to categories of decision – such as those that may impact Article 2 ECHR (right to life) or the fundamentals of a fair trial – that would not benefit from 'experimental' modelling or presumptions of proportionality and indeed which should be excluded from the purview of algorithmic tools altogether.

An accountability model for algorithmic intelligence

Algorithms in Policing – Take ALGO-CARE™ *A proposed decision-making framework for the deployment of algorithmic assessment tools in the policing context*		
A	**Advisory**	Is the assessment made by the algorithm used in an advisory capacity? Does a human officer retain decision-making discretion? What other decision-making by human officers will add objectivity to the decisions (partly) based on the algorithm?
L	**Lawful**	On a case-by-case basis, what is the policing purpose justifying the use of algorithm, both its means and ends?[1] Is the potential interference with the privacy of individuals necessary and proportionate for legitimate policing purposes? In what way will the tool improve the current system and is this demonstrable? Are the data processed by the algorithm lawfully obtained, processed and retained, according to a genuine necessity with a rational connection to a policing aim? Is the operation of the tool compliant with national guidance?
G	**Granularity**	Does the algorithm make suggestions at a sufficient level of detail/granularity, given the purpose of the algorithm and the nature of the data processed? Is data categorised to avoid 'broad-brush' grouping and results, and therefore issues potential bias? Do the benefits outweigh any technological or data quality uncertainties or gaps? Is the provenance and quality of the data sufficiently sound? Consider how often the data should be refreshed. If the tool takes a precautionary approach towards false negatives, consider the justifications for this.
O	**Ownership**	Who owns the algorithm and the data analysed? Does the force need rights to access, use and amend the source code and data analysed? How will the tool be maintained and updated? Are there any contractual or other restrictions which might limit accountability or evaluation? How is the operation of the algorithm kept secure?
C	**Challengeable**	What are the post-implementation oversight and audit mechanisms e.g. to identify any bias? Where an algorithmic tool informs criminal justice disposals, how are individuals notified of its use (as appropriate in the context of the tool's operation and purpose)?
A	**Accuracy**	Does the specification match the policing aim and decision policy? Can the stated accuracy of the algorithm be validated reasonably periodically? Can the percentage of false positives/negatives be justified? How was this method chosen as opposed to other available methods? What are the consequences of inaccurate forecasts? Does this represent an acceptable risk (in terms of both likelihood and impact)? Is the algorithmic tool deployed by those with appropriate expertise?
R	**Responsible**	Would the operation of the algorithm be considered fair? Is the use of the algorithm transparent (taking account of the context of its use), accountable and placed under review alongside other IT developments in policing? Would it be considered to be for the public interest and ethical?
E	**Explainable**	Is appropriate information available about the decision-making rule(s) and the impact that each factor has on the final score or outcome (in a similar way to a gravity matrix)? Is the force able to access and deploy a data science expert to explain and justify the algorithmic tool (in a similar way to an expert forensic pathologist)?