2017 was an extraordinary year for artificial intelligence (AI). Google developed a machine learning system that defeated the world champion at Go, learning in 40 days solutions to problems that took human players thousands of years to work out. Scientists at the Georgia Institute of Technology created a four-armed, marimba playing robot that writes its own musical compositions using deep neural networks. Researchers and film directors teamed up to produce a short film written entirely by an algorithm that named itself Benjamin. At the University of North Carolina, researchers developed a deep learning algorithm that can predict the risk of autism in babies with 81% accuracy, while at Stanford University, Google image recognition software was trained to detect and identify skin cancer with an accuracy of at least 91%.

The HART algorithm was trained using five years of custody data, totalling around 104,000 custody events.

And in the law enforcement world, Durham Constabulary became the first police force in the UK to implement a machine learning system to assess the risk of individuals re-offending. Algorithmic tools have been in use for several years elsewhere in the criminal justice system, such as the Offender Assessment System, a statistical model used by the Ministry of Justice for predicting re-offending risk. However, this is the first time that machine learning methods have been applied to such models in order to aid police decision-making, representing a major development in the police’s use of data.

Assessing re-offending risk is a core responsibility of police custody officers, as this judgement determines what further action will be taken against an arrested individual. At present, these decisions are based on a clinical model of forecasting risk, relying on the professional judgement of the human officer. Incorporating statistical methods into the risk assessment process has the potential to improve the accuracy of these judgements: research across a wide range of fields has consistently shown that statistical forecasting is typically more accurate in predicting future risk than clinical judgement alone.

According to statistics from the Ministry of Justice, in 2016 the overall re-offending rate in England and Wales was 29.6% (42.3% for juveniles). As repeat offending accounts for a large proportion of all crime, developing a system to more accurately identify those most likely to offend again in the future would allow police forces to consistently make appropriate custody decisions, resulting in a tangible reduction in overall crime rates.

The Harm Assessment Risk Tool

Durham Constabulary’s Harm Assessment Risk Tool (HART) predicts an individual’s risk of re-offending based on 34 variables, most of which relate to past criminal history. The model uses random forest forecasting (a common machine learning technique) to assess the likelihood of an individual committing a serious offence, non-serious offence, or no offence over the following 24 months. Based on this prediction, individuals are then classified into high, moderate or low risk groups.

If particular minorities have been disproportionately targeted by police action in the past, the algorithm may incorrectly assess those individuals as posing an increased risk in the future.

Machine learning-based statistical methods are able to take account of the vast amount of offender data held by police forces – far more information than a human custody officer could ever process. The HART algorithm, for instance, was trained using five years of custody data, totalling around 104,000 custody events. Methods such as random forest forecasting can balance the differential costs of errors in a way that traditional forecasting methods do not, and are therefore particularly desirable in a police setting, where the
consequences of errors can be very serious.

According to written evidence submitted to the House of Commons Science and Technology Committee, the HART model is pre-disposed to favour false positives over false negatives – it is more likely to incorrectly classify a low-risk individual as posing a high risk, as this error is considered less harmful than incorrectly classifying high-risk individuals as posing a low risk.

HART’s developers stress that the algorithm is used only to support the decision-making process of human custody officers – the officer retains ultimate responsibility for deciding what further action should be taken on the basis of the assessment. Machines will not be replacing police officers any time soon, at least not in the UK.

However, the implementation of algorithmic risk assessment tools for policing has far-reaching legal and ethical implications, particularly when we consider how this technology is likely to develop in the coming years. Of particular concern in this regard are the risk of bias, the transparency of the decision-making process, and the accountability of the decision-maker.

Accuracy and Bias

Perhaps the most common criticism of algorithmic risk assessment tools is the risk of biased outcomes. Systems underpinned by machine learning will inevitably reproduce the inherent biases present in the data they are provided with, producing predictions that reinforce or even amplify improper data recording practices. For instance, if particular minorities have been disproportionately targeted by police action in the past, the algorithm may incorrectly assess those individuals as posing an increased risk in the future. Acting on the predictions may then cause those individuals to again be disproportionately targeted by police action, creating a feedback loop whereby the predicted outcome simply becomes a self-fulfilling prophecy.

In the US, the algorithmic risk assessment tool COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) has been in use for the past 20 years, during which time it has been used to assess more than one million offenders. The system predicts an individual’s risk of re-offending within two years of assessment based on 137 variables, most of which relate to the individual’s past criminal history. Serious concerns have been raised over the use of COMPAS for decision-making. In 2016 a ProPublica investigation found that only 20% of individuals identified as likely to commit
a violent crime actually did so, and that black defendants were almost twice as likely to be deemed at risk of offending than white defendants, though this analysis is disputed by the system's developer, Northpointe.

The reasons for these apparently racist outcomes are complex. The model does not include race as a predictor variable. It does, however, include postcode. As certain postcodes have significantly higher proportions of ethnic minority residents than others, geographical bias equates to racial bias. If the data fed into the algorithm contains a disproportionately large number of arrests from a small number of 'problem neighbourhoods', the algorithm inevitably predicts that individuals from those neighbourhoods are more likely to commit crime in the future.

For many, the fact that these systems can reinforce racial bias is reason enough to preclude their use for policing and criminal justice

In reality, however, individuals from these neighbourhoods may not be any more likely to actually commit crime – just more likely to be arrested. This is the fundamental limitation of using arrest data to predict criminal activity – arrests do not directly correspond to crime incidence.

With this limitation in mind, the removal of postcode as a predictor variable may go some way to mitigating the risk of systematic racial bias. At the same time, however, if individuals from certain postcodes are in fact more likely to re-offend than others, the removal of this data will reduce the model's accuracy, which risks making the predictions less fair overall. The developers therefore face an ethical dilemma: either eliminate the variables that contribute to producing biased outcomes at the cost of reducing the model's overall predictive accuracy, or maintain a higher degree of predictive power, while conceding that a degree of bias is inevitable.

This choice may ultimately come down to legality. In the UK, the 2010 Equality Act and Article 14 of the 1998 Human Rights Act make it illegal to discriminate against an individual on a wide range of grounds, including sex, race, religion and sexual orientation. If it can be demonstrated that policing algorithms produce outcomes that disfavour certain groups of individuals on the grounds of race (even by proxy of postcode), the developers will have no choice but to remove all variables that contribute to this bias.

For many, the fact that such systems can reinforce racial bias is reason enough to preclude their use for policing and criminal justice. But rather than dismiss the algorithm outright, a more appropriate response may be to challenge the practices underpinning data collection. After all, the algorithm is not the source of the bias. It is the data that is biased, and the algorithm simply reflects the inherent biases in the dataset. These new capabilities could therefore provide insights into the limitations of existing policing practices, shining a light on potentially discriminatory methods, which would encourage police forces to re-assess whether their actions may be inadvertently targeting minority groups.

Perhaps the concern, therefore, should not be the risk of biased algorithms leading police officers to make inappropriate decisions, but rather the risk of inappropriate policing practices leading algorithms to make biased predictions.

Transparency and the Fear of the ‘Black Box’

Another ethical challenge posed by AI systems is the opacity of their decision-making process. Machine learning tools are often referred to as 'black boxes' that digest mammoth volumes of data, far beyond the capacity of a human analyst, and produce an outcome without being able to show their working. The algorithm is unable to defend its decision-making process to a human observer, as machine learning systems learn new rules through experience, rather than being directly programmed by an individual. While it is possible to design transparent AI systems, they are inevitably less powerful than their opaque counterparts, as the requirement of transparency restricts the machine's freedom to learn and evolve, significantly constraining its effectiveness and analytical 'horse-power'.

While there is a large body of research exploring the issue of bias in human decision-making, far less attention has been paid to an equally problematic influence: noise

As law enforcement agencies make increased use of (semi-)automated analysis tools underpinned by machine learning, this lack of transparency could present significant legal challenges. In order for an individual to assess whether they have been the subject of discrimination or bias in the criminal justice system, they must be able to review and scrutinise the decision-making process that led to certain actions being taken. In relation to custody decisions, the Authorised Professional Practice issued by the College of Policing dictates that the decision as to whether to grant bail to an individual or detain them in custody prior to being charged 'must be capable of withstanding scrutiny, having due regard for any supporting evidence to justify its legality, proportionality and necessity in the circumstances'. But if that decision is based on the predictions of an opaque algorithm, how can risk of bias be assessed?

In September 2017, a report by Labour MP David Lammy found that 'BAME [Black, Asian and minority ethnic] individuals still face bias, including overt discrimination, in parts of the justice system'. Addressing these inequalities will be a main priority for policymakers and practitioners across the policing and criminal justice system in the coming years. A key aspect of these reforms must be
to increase the transparency of the decision-making process, so charged
individuals are able to comprehensively assess the evidence supporting the
decisions made against them, and so decision-makers themselves can
demonstrate that their actions are legal,
fair and just. Increased reliance on black
box technology risks making the justice
process even more inscrutable, at worst
allowing discrimination and bias to persist, shrouded in a veil of ones and
zeros.

But there is a broader issue that is
often overlooked. If we are to concern
ourselves with the transparency of the
decision-making process, is a black box
algorithm any more or less transparent
than a human decision-maker? While
there is a large body of research exploring the issue of bias in human
decision-making, far less attention has been paid to an equally problematic
influence: noise.

Human judgement is strongly
influenced by a multitude of irrelevant
factors, such as mood, the weather,
time of day, fluctuations in attention,
and the stochastic nature of neuronal
firing. The result of this random noise is that if individuals are required
to make the same decision multiple
times, they will produce a number of
different judgements, and the average
of these answers will be the most
accurate. The algorithm’s decision, on the other hand, will always be the
same. While machine decision-makers
may be subject to bias, they will always
be free from noise.

So, should we be concerned by
black box decision-making? Absolutely.
But what remains unclear is whether
machines are any less transparent than
counterparts. Human judgement is affected by a range of
invisible factors that the decision-
maker is unable to fully explain when
scorned. All decision-making –
human or otherwise – includes both
transparent and opaque reasoning.

Perhaps the question, then, should
not be whether a particular prediction
by a multitude of irrelevant (guilty
March 2018, Vol. 38, No. 2
meanings in practice is designing an AI
can be fully explained, but whether it
can be sufficiently justified. What this
means in practice is designing an AI
system that is just transparent enough,
to maintain a high degree of predictive
power while still being able to provide

Accountability

Further issues arise when it becomes
necessary to hold individuals to
account regarding decisions that have
been made with the help of AI.

As systems such as HART are
intended to be used purely as advisory
tools, the custody officer retains
ultimate responsibility for ensuring that
the decisions they make are fair and
unbiased. However, regardless of who
decides what further action should be
taken, the algorithm has played a key
part of that decision-making process,
and in many cases will have caused the
officer to make a decision different to
the one they would have made without
the help of a machine.

The human users must be adequately trained to
be able to detect input errors, misinterpretation of data, and other potentially erroneous judgements made by the system

In cases such as this, the algorithm
is ultimately responsible for whatever
the final outcome is, but the human
decision-maker is the one who is held to
account. One can conceive of a situation
in which an individual has been treated
unfairly in the criminal justice system and is able to put forward a legitimate
case for discrimination. A custody
officer acting upon the predictions of
an algorithm may then face legal action
for making an unfair decision that they
would not have made had they been
acting purely on their own professional
judgement.

Of all the ethical dilemmas of AI-supported decision-making, this
is perhaps the most concerning, as there is no clear solution. Machines
cannot be held accountable for their
decisions. They cannot be found guilty of committing a crime, as criminal law
requires that accused individuals are
culpable for their actions: establishing
criminal intent, or mens rea (guilty
mind), is a necessary requirement
for proving guilt. For this reason,
accountability must rest with the human
decision-maker.

Custody officers will likely not
be aware of the legal complexities concerning accountability and AI-supported decision-making. If the
individuals using the technology do not
realise the potential worst case legal
repercussions, one should question
whether it is responsible to provide
them with such tools in the first place.

At the very least, the human users must
be adequately trained to be able to
detect input errors, misinterpretation of data, and other potentially erroneous
judgements made by the system. They
must also be made aware of the fact
that the use of AI decision support
does not in any way shift the burden of
decision-making accountability.

The advent of AI forecasting
tools undoubtedly has the potential
to enhance certain aspects of police
decision-making. But these capabilities
bring with them a range of complex
ethical and legal considerations, that
are at present quite poorly understood.

Technological progress has a tendency
to outpace legislative change, and
government must carefully consider
whether new policy is needed to allow
police forces to take full advantage
of these new tools without fear of
compromising their legal and ethical
responsibilities.

Alexander Babuta
Alexander is a Research Fellow in the National Security and Resilience Studies
group at RUSI. His research focuses
on policing and security in the digital
decade, transnational organised crime and
counterterrorism. He holds an MSc
with Distinction in Crime Science from
University College London (UCL),
where his dissertation explored the
use of data science methods for risk
assessing missing children.

The views expressed in this article are those of
the author, and do not necessarily reflect those
of RUSI or any other institution.