The constitutional dangers of algorithmic decision-making systems in the criminal justice domain

The Norwegian Association for Computers and Law
Knut Selmer Annual Memorial Lecture

23rd November 2020

Professor Karen Yeung

University of Birmingham
Law School & School of Computer Science

Introduction

- On-going 4 year VW Stiftung funded project on ADM in criminal justice, from law, computer science, political science and neuropsychology, with Dr Adam Harkens (University of Birmingham)

- Rapidly growing use of ADM systems in the public sector, including the criminal justice domain

- Controversial – esp concerns about racial bias and unfair discrimination

- Today: ADM tools and systems developed through machine learning (ML) techniques that purport to evaluate the ‘risks’ posed by individuals

Principal message

- Highlight tendency in current approaches to the design, of ADM systems to ‘decouple’ the algorithmic model from its context, thereby failing to take constitutional principles seriously, creating risks of arbitrary and unjust decisions

- To nurture and sustain our communities as liberal democratic polities committed to respect for individual freedom, democracy and the rule of law, algorithmic systems intended for use in the criminal justice context must pass constitutional muster otherwise they cannot be justified and should not be used at all

Outline & structure

1. Why the criminal justice domain is unique, powerful and dangerous

2. Three examples of ADM systems that utilise machine learning (ML) to evaluate or identify ‘risky’ individuals (SyRI, London Gangs Matrix, HART tool)

3. Highlight current approaches to the design, implementation and evaluation of ADM systems pose constitutional dangers due to the ‘decoupling**’** of ADM tools from vital elements of the context in which they are used

4. Illustrate via the Durham Constabulary’s HART forecasting tool

1. Conclusion

1. Criminal justice and constitutional safeguards

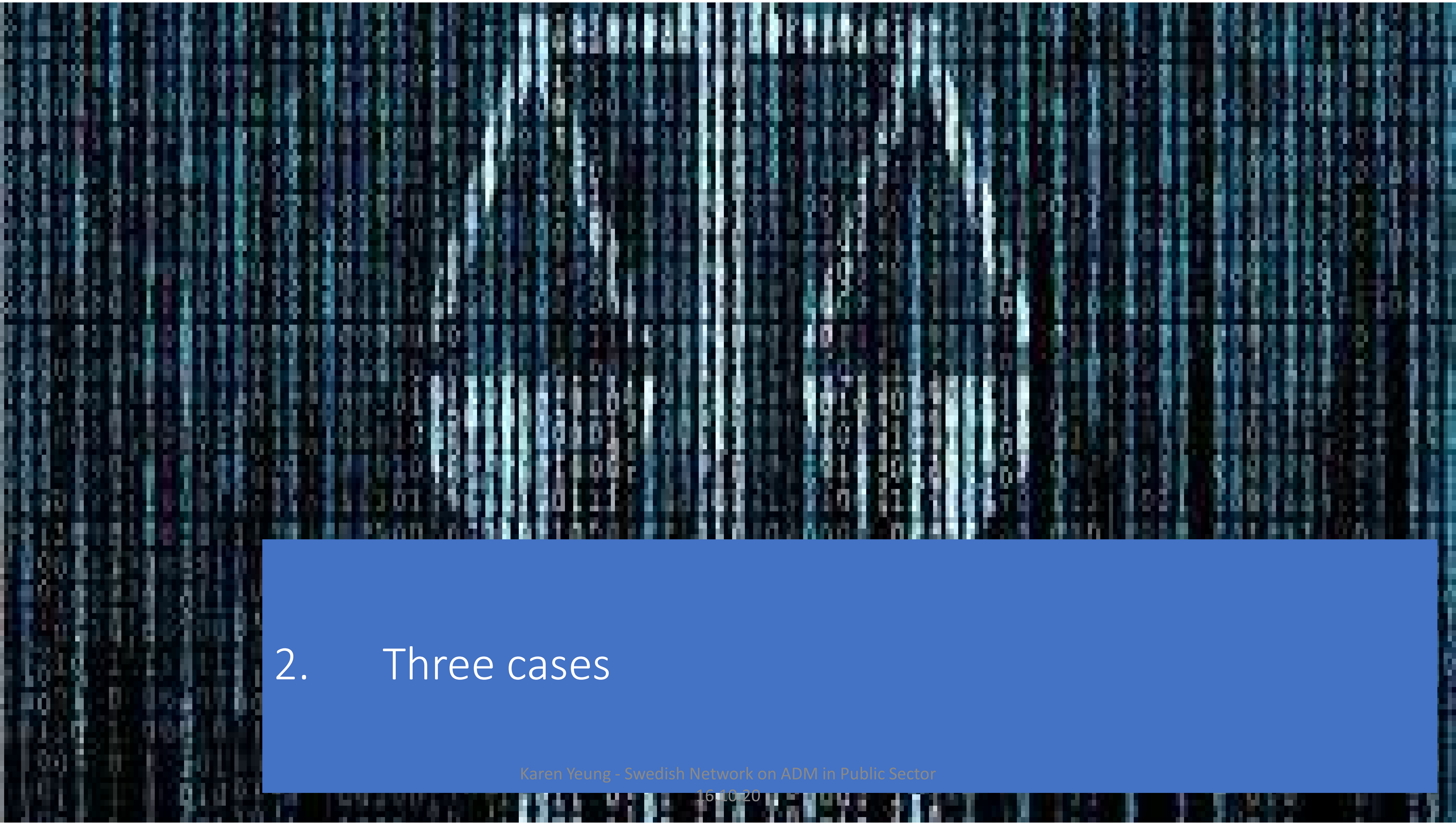
Why makes the criminal justice domain special?

- State's monopoly on the exercise of **legitimate coercion** is at its most vivid and powerful.

- Subject individuals to **state-sanctioned punishment** for the commission of criminal offences, resulting in serious restrictions on individual liberty and/or the deprivation of property accompanied by significant moral and social stigma (beyond that which would be possible in a civil or administrative action).

- Constitutional constraints are therefore vital**,** pointing to the importance of institutional and legal ***safeguards against the abuse of power*** which is the ultimate aim of a commitment to constitutionalism

- Argue: use of ML techniques to create ADM tools to inform CJ decisions introduce new forms of arbitrariness and expand the scope for the abuse of governmental power



2. Three cases

Case studies

- **SyRI** (Government of the Netherlands)
- **Gangs Matrix** (London Metropolitan Police)
- **HART** (Durham Constabulary)

Case study 1: SyRI

- ML-based decision-support system used to help detect, prevent and combat fraud against the government by the Dutch Government, introduced in 2008

- Aggregation of individual records from multiple government agencies, subject to SyRI to identify ‘suspect’ individuals ie those having ‘an **increased risk of irregularities**’ when compared to the rest of the target population – indicating the ‘unlawful use of government funds.’ In-depth criminal investigation may follow.

- Hague District Court held unlawful (February 2020) because disproportionate interference with Art.8 ECHR (Right to Privacy).

Case study 2: London Gangs Matrix

- The Gangs Matrix was created by the London Metropolitan Police Service aftermath of the London Riots in August 2011 to reduce gang violence

- The Gangs Matrix is produced by police in cooperation with partner agencies to create a shared digital dashboard listing individuals identified as a ’potential gang member’ (gang nominal) either by police or representative from partner agency based on ‘reliable evidence from more than one source

- Potential ‘victims’ of gang violence can also be included in the Matrix

London Gangs Matrix

- ML model is used to evaluate each individual's 'harm score' (individual risk 'to others'): categorised either as **red** (high risk of violent offending), **amber** (medium risk of violent offending), or **green** (low risk of violent offending). These 'harm scores' are accessible at any time by front-line police officers. Potential victims are not given a harm score.

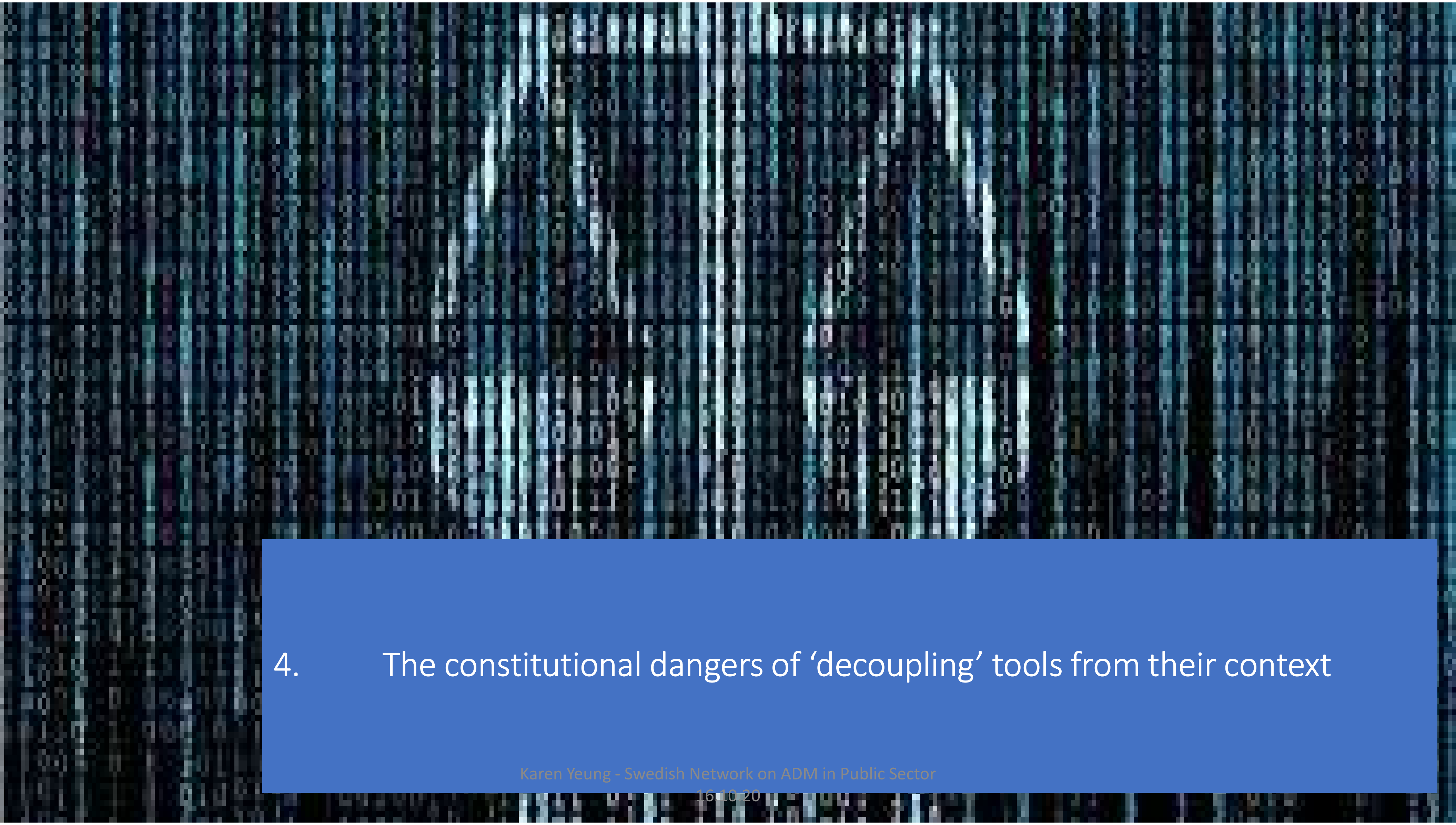
- Factors that contribute to 'harm score' have been made public, but algorithmic is not available so we do not know how the harm scores are generated

- Gangs Matrix used by police and partner agencies to inform 'graded response' approach, based on harm scoring intended to be 'commensurate with risk.'

- **Policing Interventions include:** Heightened police surveillance for specific local areas and individuals, often leading to pre-emptive stop and search, and arrest for minor offences;

Case study 3 – HART Tool

- Durham Constabulary's **Checkpoint programme began in 2015**, "with the aim of reducing the number of victims of crime through an innovative approach to cut re-offending by offering alternatives to prosecution, to certain identified individuals.
- The Harm Assessment Risk Tool (HART) is used to inform these decisions about whether to offer suspects an opportunity to participate in Checkpoint by assessing assessment of the suspect being arrested for suspected (violent) crime in the next 2 years
- Developed by statisticians from University of Cambridge in collaboration with Durham police.
- Durham Constabulary's Head of Criminal Justice, Sheena Urwin, sought to investigate the quality of HART predictions, via Master's thesis: published on-line. So, more transparency than typical, although algorithmic model not published



4. The constitutional dangers of ‘decoupling’ tools from their context

3. The constitutional dangers of ‘detachment’

- Technical developers who design ADM tools for public sector use prone to ‘detach’ the mathematical tool from their application context

- Three ‘decoupling’ practices, detaching the algorithmic decision-tool from :
  - **substantive intervention** that follows from the recommendation generated by the tool.
  - **policy purposes** which the tool is intended to serve.
  - **impact** of the tool on affected individuals, others and society more generally.

- This leads to the side-lining of foundational constitutional principles, thereby generating substantial risks of arbitrariness and injustice

Decoupling the tool from the substantive intervention

(a) The problem of mistakes:

- **Question:** How are the ***risks*** of Type I (false positive) and Type II (false negative) errors distributed by the ADM tool itself?

- **Context matters:** The mistakes of a recommendation tool (*e.g.* Amazon shopping) are significantly and substantively different to those in the CJ context.

- **Perspective matters:** Which mistakes are 'better'? The answer this depends on whose perspective is under consideration (*e.g.* in criminal justice: the police? the public? the individual being evaluated?

- ADM systems require the design of 'error' thresholds so as to attempt to accurately profile and categorise individuals.

- Respect for the constitutional consequences of mistakes *must* be hardwired into the design of any ADM tool in the criminal justice context.

(1) Decoupling the tool from the substantive intervention

(a) Distribution of Type I and Type II error

- Criminal justice system founded on presumption of innocence

- Type I error (falsely convicting the innocent) much more serious moral harm than Type II error (letting the guilty go free) although latter is deeply regrettable

- Right to due process requires minimization of Type I error, although might increase risk of Type II error

- Yet no apparent awareness of this by ADM technical developers

Decoupling the tool from the substantive intervention

(b) Procedural dimensions

What, if anything, should an affected individual be informed of when subject to the evaluation of a ADM system and should they have an opportunity to challenge and contest?

- **Context matters:** Would these answers differ when discussing an Amazon recommendation vs. a criminal justice decision?

- **Technical features matter:** ML-based systems lack the explanatory power of rule-based (and to a lesser extent) statistical tools, because of **(a)** their **inscrutability**, and **(b)** the non-intuitive outputs which they generate.

Mathematical models and reasoned explanations

i. Rule-based systems:

- Operate on an 'if *x* then *y'* basis (*E.g.* automated parking ticketing systems: has a car parked in this prohibited parking space? *Yes*).

- Need not be digital and the rules for each systems are pre-specified.

ii. Statistical systems:

- Can incorporate more complex information than rule-based systems allowing the production of 'risk scores' comparing an individuals with wider populations (e.g. Offender management systems: combination of interview data with criminal history etc.).

- While statistical systems may not be causally explainable, they are at least functionally explainable (may be possible to demonstrate *plausible* relationships for generated outputs).

Iii ML-based systems:

- Capable of handling more complex data, at a much greater volume than the other systems. In doing so, they sacrifice the capacity of decision-makers to meaningfully question the underlying causal relationship between data inputs and generated risk profiles.

- Unlike statistical tools, ML systems are additionally ***not*** functionally explainable, meaning that outputs may be produced without ***any*** intuitive explanation being made to the public official tasked with making decisions on the basis of the system's recommendation.

Decoupling the tool from the substantive intervention

(c) Responsibilisation and the 'seeing is solving' fallacy

- Beware the 'digital enchantment' reflected in celebratory rhetoric dominating contemporary discussions of the 'need and urgency' for the public sector to embrace automation.

- Predictive systems in CJ domain tend to be used to target 'risky individuals' in order to motivate or otherwise incentivise them to change their behaviour (*i.e.* **responsibilisation**).

- **'Seeing is solving' fallacy:** though data enables us to 'see' or 'visualise' the nature and dimensions of a problem, it tells us nothing about what we ***should do*** in response to the problem.

Decoupling the tool from the substantive intervention

The ‘seeing is solving’ fallacy

- The challenges of the CJ domain are **complex, sensitive and highly contextual**, meaning that any substantive intervention flowing from the output generated by an ADM system must reflect this
- **London Gangs Matrix**: intensification of surveillance and monitoring for ’gang nominals’. But is fostering a culture of suspicion, entailing the intensification of surveillance and warnings targeted at children identified as vulnerable to gang really ‘better’?
- **Unwin** thesis 2016 (HART tool):

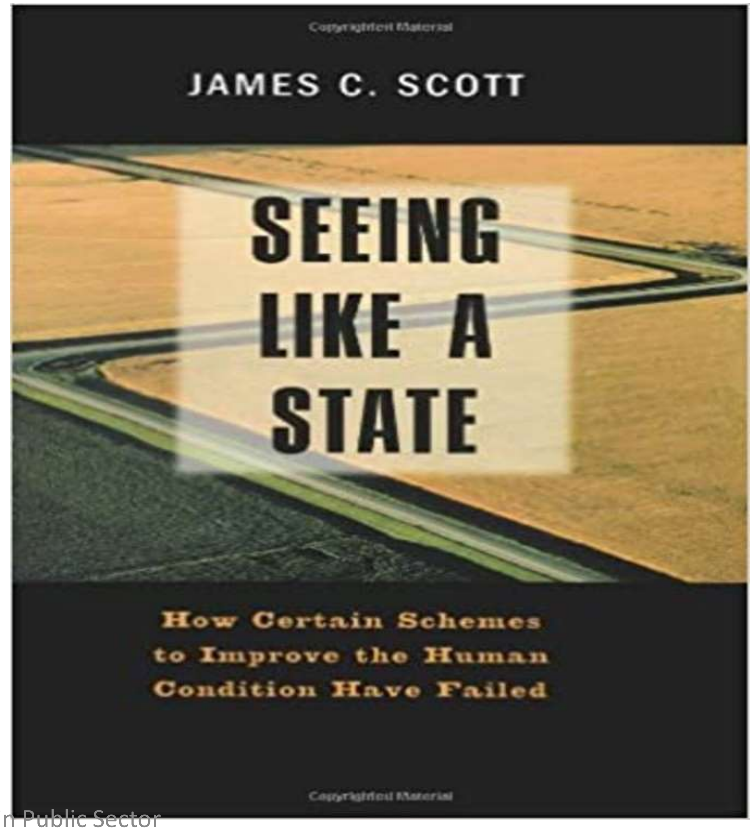
“While ML is a developing field, and multiple different approaches continue to compete with one another, one certain conclusion can be drawn from this literature: once a prediction is made, **the accuracy of the prediction is what determines its success”** (p 25)

The hollow promise of ACCURATE prediction

- Beware the hollow promise of allegedly ‘accurate’ risk predictions esp about individuals . Even *assuming* these predictions are **accurate**, that may tell us where to look, but little about *what we should do* (ie the nature of the appropriate substantive intervention).

- Recall **James C Scott’s** *Seeing Like a State:*

The detailed maps of Amsterdam that identified the location of Jewish families could just as easily be used to provide those families with food and care as much as to round them up and send them to Concentration camps.



(2). Decoupling the tool from the policy purpose

The tendency to ‘***decouple’*** the underlying policy purposes for which an ADM tool is intended to promote from the process of design and implementation.

Results in tendency to ignore:

a) Does the tool have a lawful basis and comply with all applicable legal requirements

b) Is the training data an accurate proxy for ‘ground truth’?

Decoupling the tool from its policy purpose

(a) Is the tool legally permissible?

- Requires attention to the underlying statutory context governing its use.

- **Recidivism risk:** Is it lawful to built a tool to inform decision whether to remand an individual on the basis of a predicted risk of committing a crime (or in fact, probability of being arrested for a suspected crime in future) of legislation requires decision to be based on assessment of 'flight risk' ?

- Consideration of lawfulness must occur **before** a tool is built yet often overlooked

Decoupling the tool from its policy purpose

(b) Is the training data an accurate proxy for 'ground truth'?

- To build an ML-based tool that generates accurate predictions requires not only very large datasets, but the data itself must be 'high quality' (reflecting 'ground truth' of targeted phenomenon).

- **Predicting crime raises two insuperable problems: (a)** impossible to identify all true positive in past data, and **(b)** one can never comprehensively identify false negatives.

- Thus, ADM systems in the CJ domain are reliant on 'proxy' data which fails to accurately reflect 'ground truth'.

- ADM systems are further quality assessed by data scientists on the basis of 'predictive accuracy', which suffers from the very same flaws. No agreed standards concerning data quality.

- Developers must take more critical, rigorous approach to questioning the integrity and provenance of data used to build predictive models.

(3) Decoupling the tool from its impacts

- The tendency for public organisations and officials (including, but not limited to, the CJ domain) **seek to downplay the adverse impact of ADM systems and tools**, by emphasising *claims* that the new tool offers 'better' service.
- By invoking this '**argument of analogy**', public organisations then argue that explicit legal authority is not required to adopt the new tool because, **they claim**, that the new ADM system is simply a better, faster, and more efficient method of decision-making that was previously undertaken.
- But different ADM systems and tools will have different side-effects, although aimed at achievement of the same goal. Hence different tolls have different impacts upon legal rights and interests
- Eg SyRI tool: population wide data-veillance a serious threat to individual freedom and autonomy (disproportionate intervention with right to privacy)

Decoupling the tool from its impacts

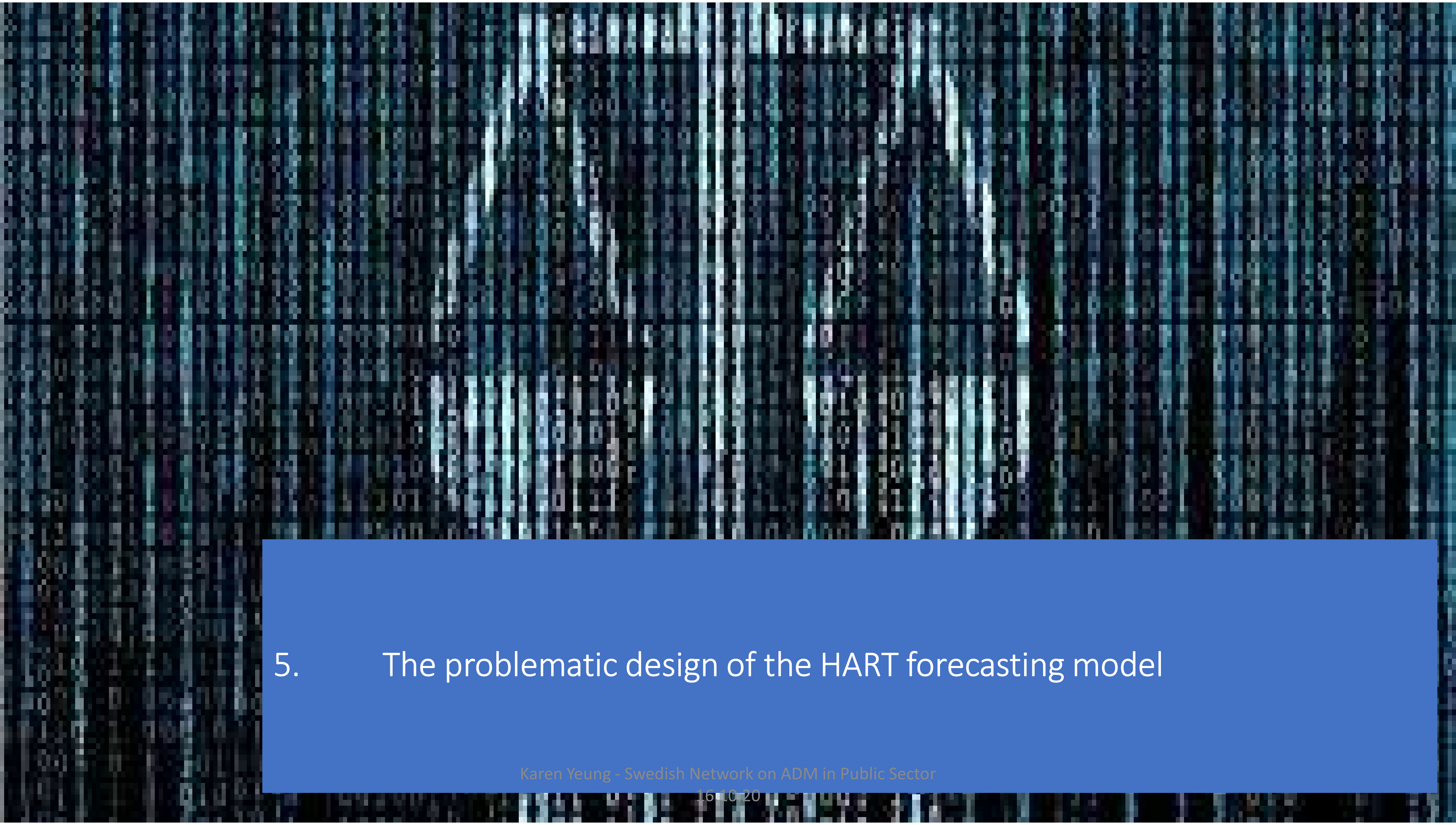
Algorithms don't harm people fallacy'

(Gun's don't harm people, people harm people..)

- Even if a tool does not, in and of itself, interfere with or engage human rights, it does not follow that we can ignore other kinds of harm: like

- Troubling tendency of 'digital enchantment' in contemporary policy discussions, to ignore that ADM systems are powerful technologies which are capable of producing dangerous decisions (even if tempered by a team of benevolent developers, or the 'perfect' exercise of discretion by a human decision-maker), because of the capacity to operate automatically, at scale, and to trigger action that is remote in both time and space from the location at which the action is triggered.

- *E.g.* 'government by database' much more threatening to human rights and democracy than a policeman with a pen and notebook



5. The problematic design of the HART forecasting model

The Durham HART tool

HART is a forecasting tool created through the application of ML techniques to arrest data held by Durham police.

Generates prediction of an individuals ‘risk of re-offending’.

This prediction is then provided to the custody officer, who retains the discretionary power to follow or reject this prediction.

- **High risk =** a suspect is deemed likely to commit a serious offence over the next two years.
- **Medium risk =** a suspect is deemed likely to commit *any* offence over the next two years.
- **Low risk =** a suspect is deemed unlikely to commit an offence over the next two years.

Only ‘medium risk’ individuals eligible for Checkpoint, but others proceed via standard criminal process

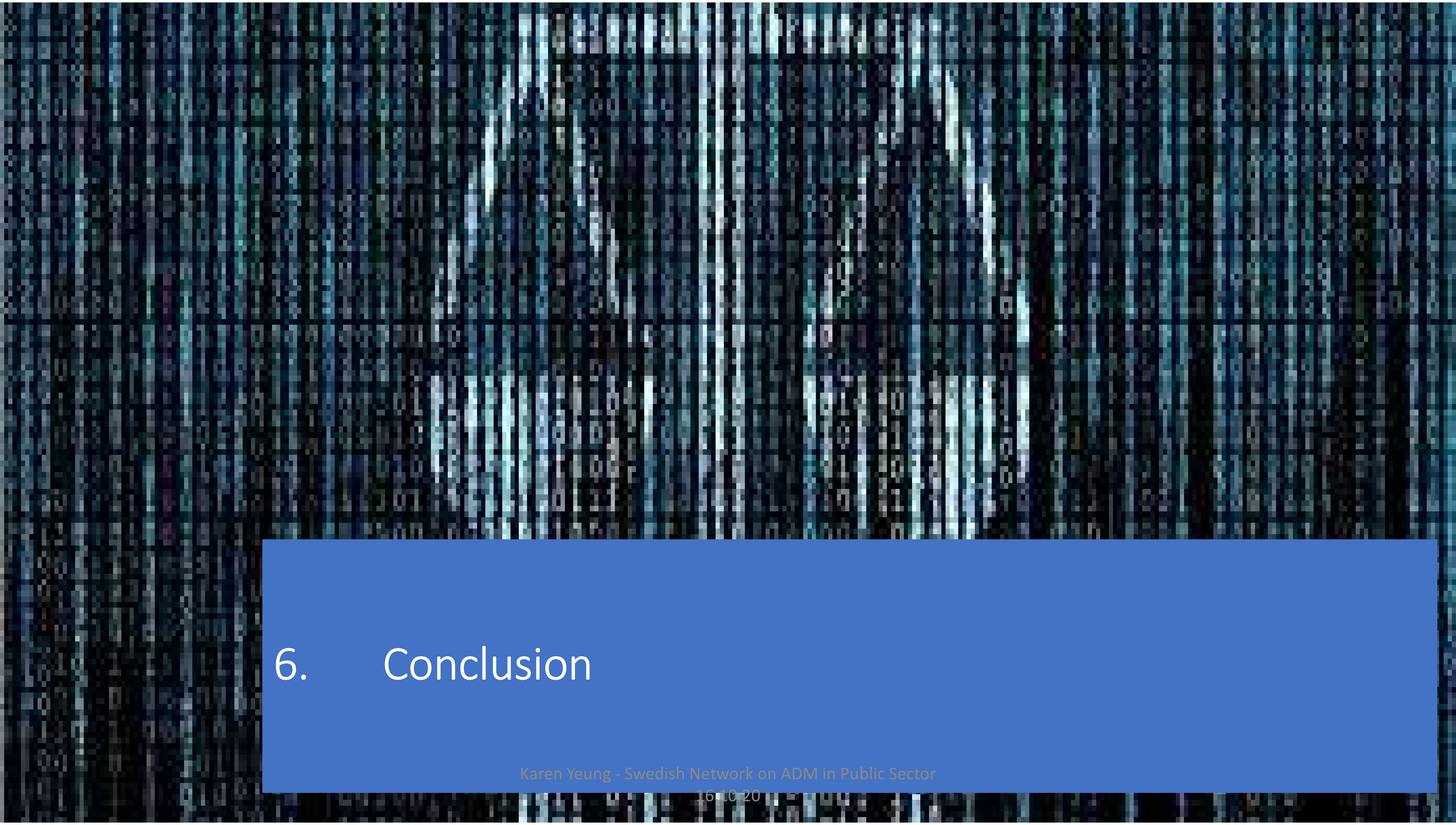
HART’s design logic

a) Uses ‘arrest’ data as proxy for crime to predict risk of ‘offending’. It is NOT an a predictor of ‘offender dangerousness’. Rather, it is a predictor of **re-arrest** for a crime in the next 2 years

b) Both ‘high risk’ and low risk individuals excluded from Checkpoint. So, if mistaken prediction that high risk individual is in fact low risk – fortuitously, results in same intervention. Ie Configuration of Type I and Type II errors have been erroneously dealt with (in light of the stated policy purpose of the ADM tool.

c) Why are ‘high risk’ individuals excluded (since they might benefit the most)?

- But, no serious injustice, because HART tool used to determine whether suspect offered a ‘benefit’ and no intervention of legal or other rights entailed.
- Would be better to describe in more modest, accurate, policy-directed terms.



6. Conclusion

Conclusion

- I have highlighted three common ‘decoupling’ practices in the construction of ADM tools for use in criminal justice contexts
- These generate several dangers that may give rise to serious and substantial injustice, due to failure to take seriously several foundational constitutional principles that are intended to operate as safeguards against the abuse of power, notably:

- Right to due process
- The rule of law
- Public law principles of relevance, proportionality and justification

Conclusion

Several areas of law are important vehicles through which ADM systems might be subject to legal challenge, including

- human rights law,
- administrative law and the principles of judicial review,
- contemporary data protection law and
- anti-discrimination legislation.

So, the law provides an important, concrete institutional mechanisms for securing algorithmic accountability.

Conclusion

- Safeguards against the abuse of algorithmic power esp important in criminal justice decision-making
- Due to inscrutability and potential to generate arbitrary outputs, use to inform decisions resulting in the deprivation of rights should be constitutionally impermissible

Conclusion

Courts and judges cannot provide sufficient constitutional protection because:

1. Litigation ex post, limited, unsuited to handle collective action problems
2. Inherent limits of existing legal rights and mechanisms
3. Woeful lack of transparency, consultation and debate about ADM in public sector

Conclusion

To design ADM systems that can add real value to public sector decision-making, need to combine and integrate technical and legal expertise

- Yet lawyers are trained to evaluate the legality by critical examination of texts, transactions and social practices, not algorithmic tools
- Law students are not provided with basic data training: perhaps in future?

University of Birmigham's new **M Sc in Responsible Data Science**: taught by Law School + School of Computer Science to lawyers (2021):

https://www.birmingham.ac.uk/rds