Risk factors and mitigation measures for AI use in and by the Public Sector

Paul Waller

Brainteaser

5% of children in a population are in danger of domestic abuse.

A predictive classification algorithm correctly identifies a child as in danger for 80% of those truly in danger.

It correctly identifies as safe 90% of those not in danger.

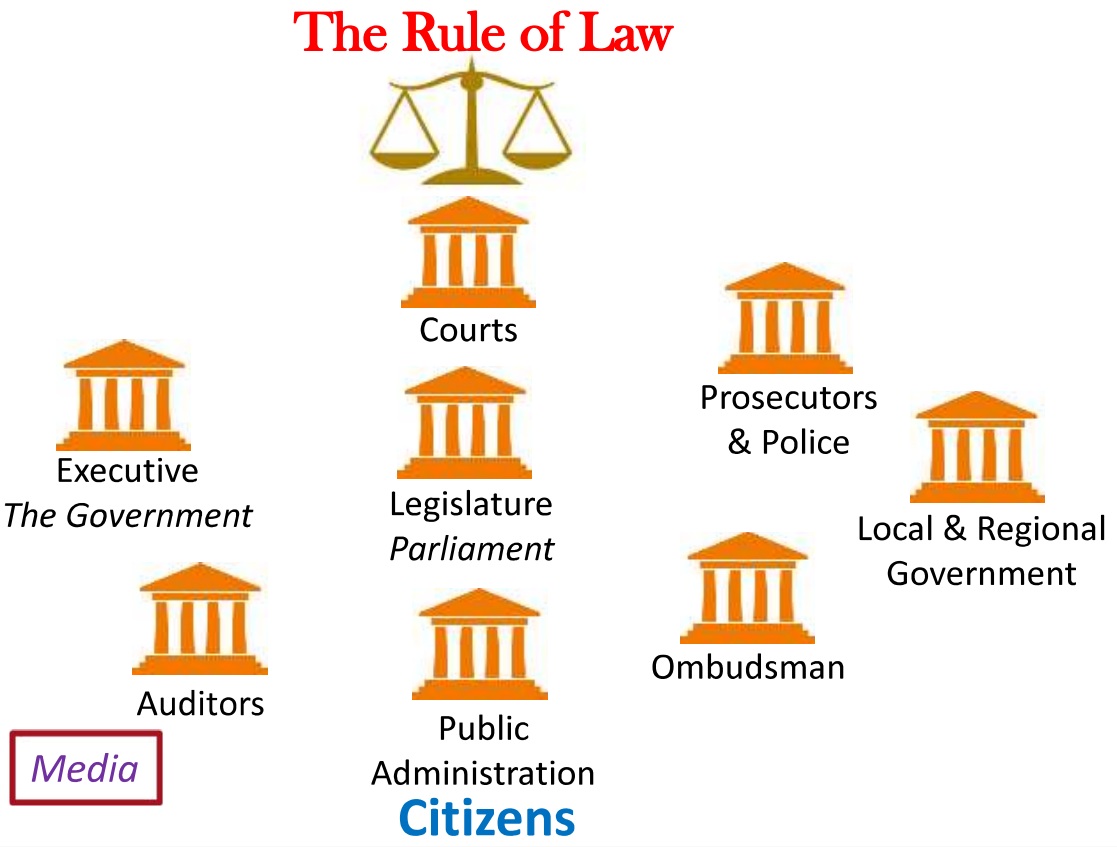
If the algorithm identifies a specific child as in danger, what is the probability that the child truly is in danger?

Approximately:

a) 90% b) 80% c) 72%

d) 30% e) 10% f) 5%

The public sector context



Bad news…

Ofqual exam results algorithm was unlawful, says Labour

Exclusive: shadow attorney general says ministers would have been aware of at least three breaches of the law

Boris Johnson's 'mutant' planning algorithm could scar England for ever

UK passport photo checker shows bias against dark-skinned women

NHS Digital reviewing algorithm after women incorrectly told to shield

Councils scrapping use of algorithms in benefit and welfare decisions

Call for more transparency on how such tools are used in public services as 20 councils stop using computer algorithms

Police built an AI to predict violent crime. It was seriously flawed

A Home Office-funded project that used artificial intelligence to predict gun and knife crime was found to be wildly inaccurate

This image-recognition roulette is all fun and games... until it labels you a rape suspect, divorcee, or a racial slur

If we could stop teaching AI insults, that would be great

It’s a risky business

Madeleine Waller and Paul Waller. *Why Predictive Algorithms are So Risky for Public Sector Bodies*, 2020. http://dx.doi.org/10.2139/ssrn.3716166

Legal Risks

- The European Convention of Human Rights (ECHR)
- The European Social Charter (ESC)
- The International Bill of Human Rights
- The Charter of Fundamental Rights of the European Union (CFR)
- General Data Protection Regulation
- Freedom of Information Acts
- Domain Specific Legal Instruments
- Legal Instruments Protecting Particular Groups
- Administrative Law (mandate) for the functions being exercised

Source: https://www.turing.ac.uk/research/publications/ai-human-rights-democracy-and-rule-law-primer-prepared-council-europe

- The European Code of Good Administrative Behaviour

Source: https://www.ombudsman.europa.eu/en/publication/en/3510

Good Administrative Behaviour

Lawfulness, clear governance and accountability,

Respect for human rights including the right to privacy,

Accuracy in relation to the public function being exercised,

Equality and consistency of treatment and absence of bias or discrimination,

Clarity of the explanations for decision making and reasons for decisions,

Absence of negative consequences,

Security,

Proper record keeping.

Data Risks

My Top Five Sources of Risk

- Bias
- Unrepresentativeness
- Quality
- Flawed data pre-processing/coding
- Invalid statistical assumptions

Reality defies datafication

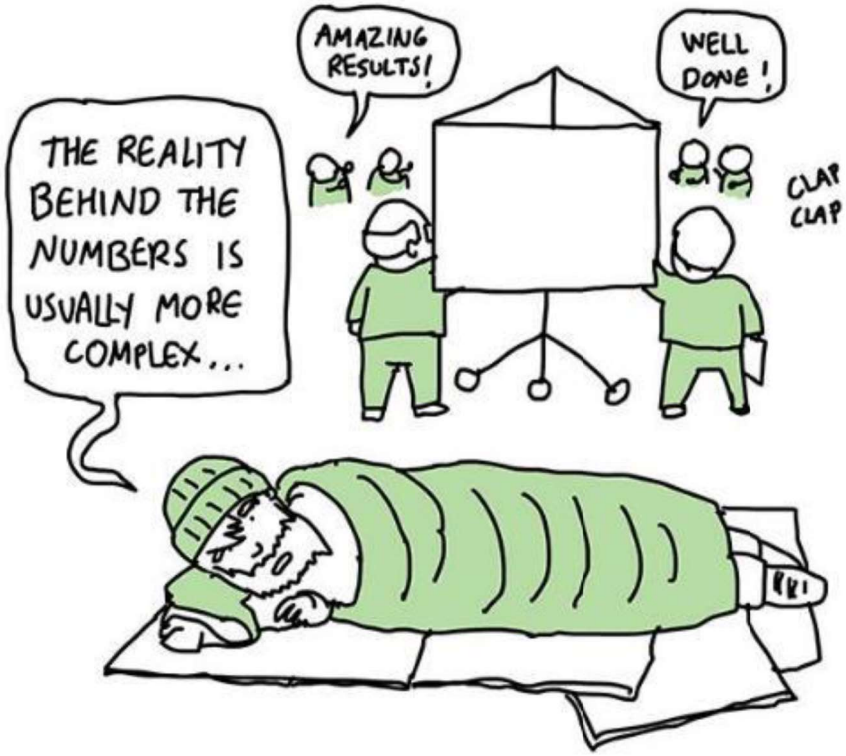


Image from Made to Measure, by Toby Lowe.

Design Risks

My Top Five Sources of Risk

- Choice of model relative to problem & data
- Specification of model & optimisation parameters
- Parameter initialisation
- Inadequate testing
- Incomprehensible complexity

Implementation Risks

My Top Five Sources of Risk

- Poor operational testing
- Inadequate security
- Poor contract management
- Inadequate process design
- Inadequate training

Use Risks

My Top Five Sources of Risk

- Inaccuracy
- Lack of understanding of probabilistic measures & ranges, and weighting consequences
- Automation bias/aversion
- Obscure or inexplicable working and outcome
- Abuse of privacy & other human rights

Use Risks

95% Safe to eat

5% You will die

Do you eat it?

95% Win the race

5% Lose the race

Do you bet €20 on a win?

Use Risks

5% of children in a population are in danger of domestic abuse.

A predictive classification algorithm correctly identifies a child as in danger for 80% of those truly in danger.

It correctly identifies as safe 90% of those not in danger.

If the algorithm identifies a specific child as in danger, what is the probability that the child truly is in danger?

Approximately:

a) 90% b) 80% c) 72%

d) 30% e) 10% f) 5%

Use Risks

5% of children in a population are in danger of domestic abuse.

A predictive classification algorithm correctly identifies a child as in danger for 80% of those truly in danger.

It correctly identifies as safe 90% of those not in danger.

If the algorithm identifies a specific child as in danger, what is the probability that the child truly is in danger?

Approximately:

a) 90% b) 80% c) 72%

d) 30% e) 10% f) 5%

Assurance

- What's the outcome we want?
- Is it lawful to do this?
- Is the data there & OK?
- Even if it works, is it wise?
- Will it work??!
- Do we understand it, can we explain it?
- Can we actually get it working well in reality?

Brainteaser - solution

5% of children in a population of 1000 are in danger of domestic abuse: 50 (so 950 are not)

A predictive classification algorithm correctly identifies a child as in danger for 80% of those truly in danger: 40 (so it misses 10)

It correctly identifies as safe 90% of those not in danger: 855 (misidentifying 95)

	Truly in danger	Not in danger	Total
Identified as in danger	40	95	135
Identified not in danger	10	855	865
Total	50	950	1000

So the algorithm identifies 135 children as in danger, of which 40 truly are in danger, giving a probability of 40/135 = 0.296 or approx 30%

But the "Accuracy" is % correct identification = (855+40)/1000 = 89.5% !!!

Risk factors and mitigation measures for AI use in and by the Public Sector

Paul Waller

Brainteaser - solution

"D" represents a child truly in danger of domestic abuse
"F" represents the test flagging a case as "in danger"
We are given (where "~" means "not"):
$P(D) = 5\% = 0.05$ so $P(\sim D) = 0.95$
$P(F \mid D) = 80\% = 0.80$
$P(\sim F \mid \sim D) = 90\% = 0.90$ so $P(F \mid \sim D) = 0.10$

We need to find $P(D \mid F)$, the probability of truly in danger if flagged

Now $P(F) \times P(D \mid F) = P(D \wedge F) = P(D) \times P(F \mid D) = 0.05 \times 0.80 = 0.04$

So $P(D \mid F) = P(D) \times P(F \mid D) / P(F) = 0.04 / P(F)$ where

$P(F) = P(F \wedge D) + P(F \wedge \sim D) = 0.04 + P(\sim D) \times P(F \mid \sim D) = 0.04 + 0.95 \times 0.10$
$= 0.04 + 0.095 = 0.135$

Therefore $P(D \mid F) = 0.04 / 0.135 = 40 / 135 = 0.296$ or approx 30%