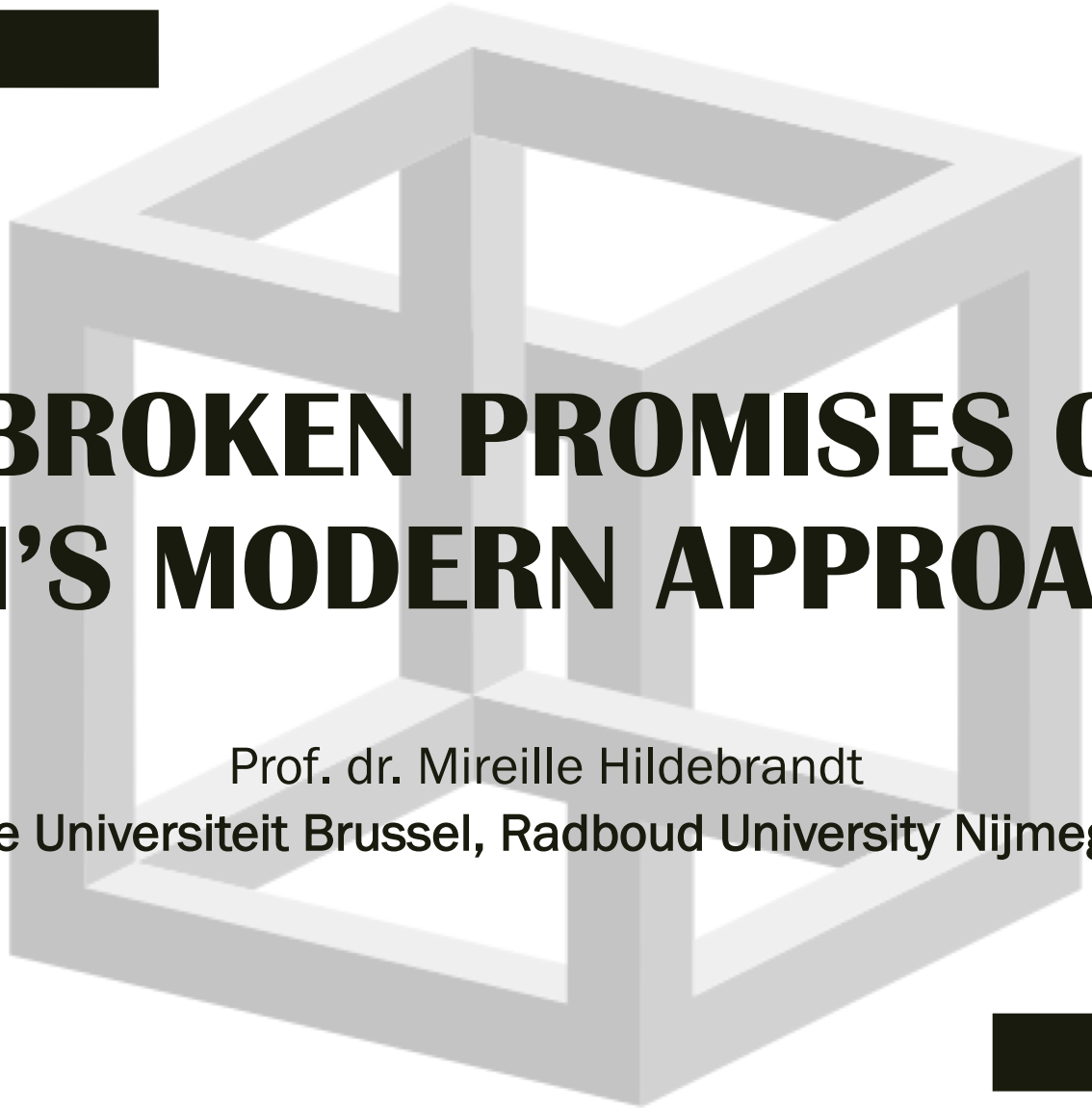


BROKEN PROMISES OF AI'S MODERN APPROACH

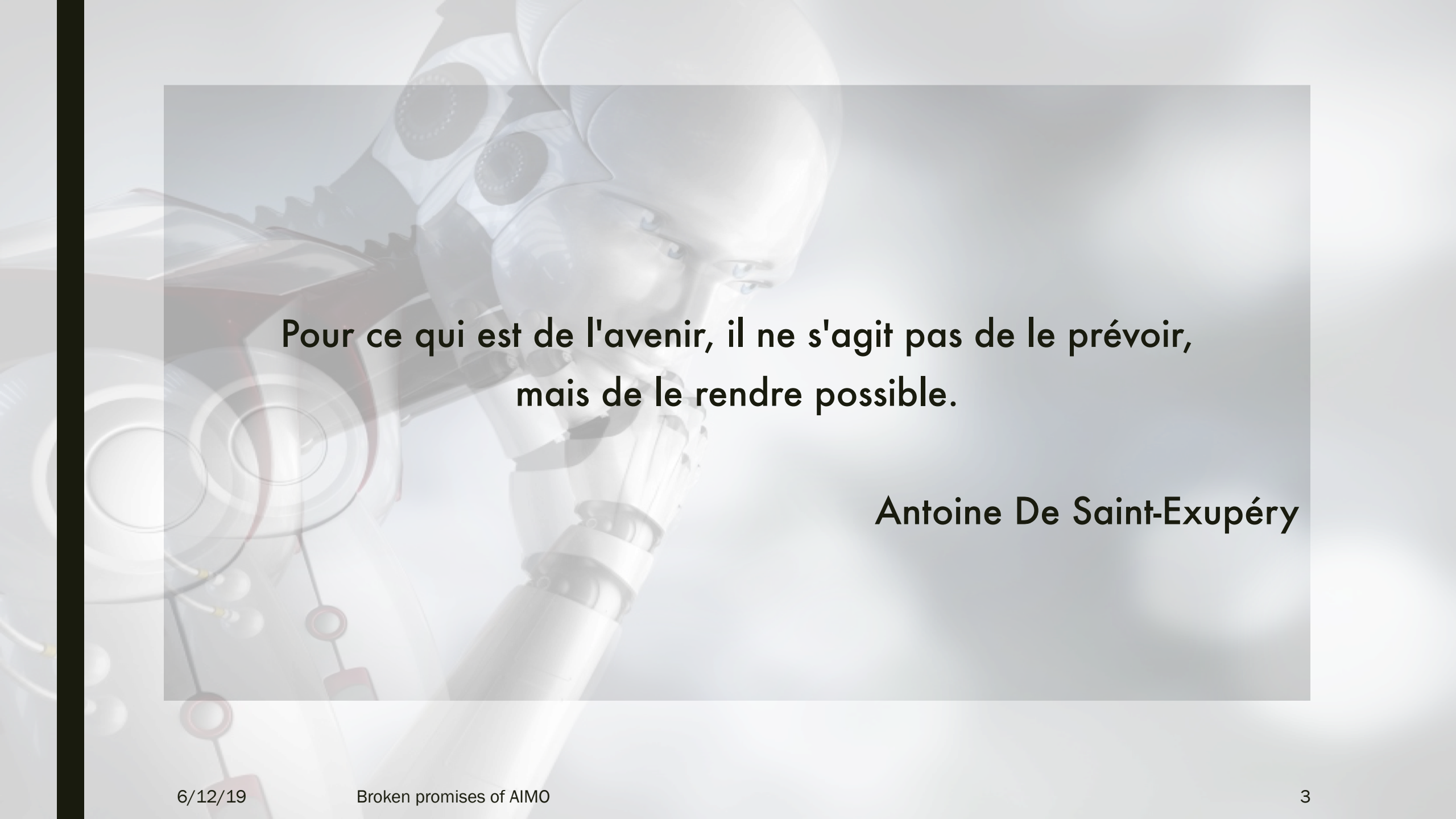
Prof. dr. Mireille Hildebrandt
Vrije Universiteit Brussel, Radboud University Nijmegen

European Research Council
Established by the European Commission

Since the present futures co-determine the future present,
predictions basically enlarge the probability space we face;
they do not reduce but expand both uncertainty and possibility.

The question is about the distribution
of the uncertainty and the possibility:
who gets how much of what?

Hildebrandt 2016



Pour ce qui est de l'avenir, il ne s'agit pas de le prévoir,
mais de le rendre possible.

Antoine De Saint-Exupéry

Law, I have been saying, somewhat against the pretensions encoded in woolsack rhetoric, is local knowledge; local not just as to place, time, class and variety of issue, but as to accent—vernacular characterizations of what happens connected to vernacular imaginings of what can. It is this complex of characterizations and imaginings, stories about events cast in imagery about principles, that I have been calling a legal sensibility. This is doubtless more than a little vague, but as Wittgenstein, the patron saint of what is going on here, remarked, a veridical picture of an indistinct object is not after all a clear one but an indistinct one. Better to paint the sea like Turner than attempt to make of it a Constable cow. (Geertz 1983, 215)

What's next?

- AIMA and the nature of prediction
- Broken promises?
- AIMA in policing
- Diminishing whose human agency?
- The 'human in the loop'?
- Rule of law and the nature of discretion

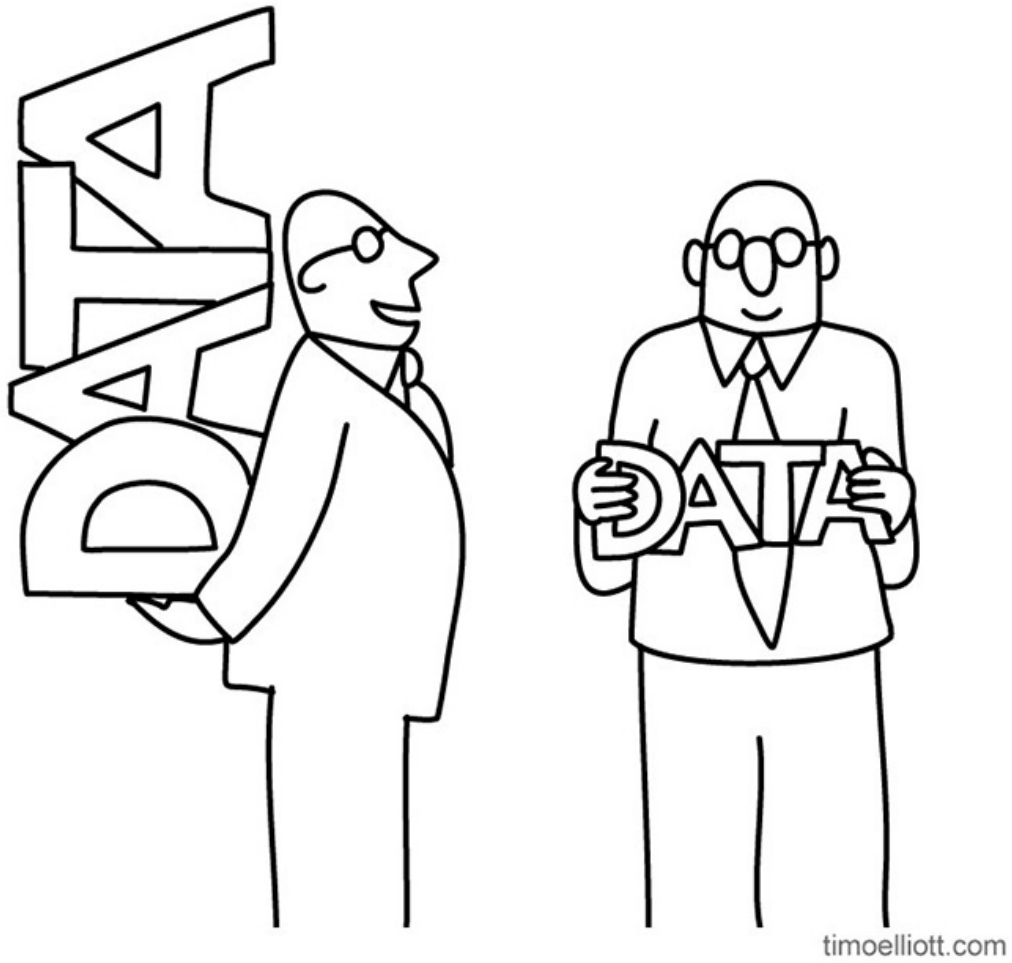
What's next?

- AIMA and the nature of prediction
- Broken promises?
- AIMA in policing
- Diminishing whose human agency?
- The 'human in the loop'?
- Rule of law and the nature of discretion

AIMA and the nature of prediction

- From code-driven to data-driven AI

- **GOFAI (code-driven)**
  - *Translating acquired knowledge into computer code*
    - Formalisation
    - Disambiguation
  - *Relationship with legal certainty and explainability*



"I think you'll find that mine is bigger..."

AIMA and the nature of prediction

- From code-driven to data-driven AI

- **AIMA (data driven)**
  - *Training a 'learner' on relevant data (training data)*
  - *Framed in terms of a feature space (if supervised)*
  - *Tested in terms of a hypothesis space (mathematical functions)*
  - *Tested against validation data (e.g. by way of cross-validation)*
  - *Verification against test data (out of sample)*
  - *Based on a machine readable task*
  - *Measured by a performance metric*

AIMA and the nature of prediction

- From code-driven to data-driven AI

- AIMA
  - *Exploratory*
  - *Confirmatory*
  - *Reality gaps*
  - *How can we check reliability?*
  - *Who is we?*

What's next?

- AIMA and the nature of prediction
- Broken promises?
- AIMA in policing
- Diminishing whose human agency?
- The 'human in the loop'?
- Rule of law and the nature of discretion

Broken promises of AIMO

- Remember the 'broken windows theory'?
  - *Kelling & Wilson 1982*
  - *Criminological theory that says 'untending' to destruction nourishes crime*
  - *They claim that reduction in 'disorderly behaviour' works*
  - *Giuliani 'proved' it works during his tenure as mayor of Manhattan*
  - *Harcourt: (1) regression to the mean, (2) instigated stop & frisk*
- Remember the 'broken window fallacy'?
  - *Frederic Bastiat 1850*
  - *Economic theory that suggests that destruction is economically beneficial*
  - *Profits on one side loss at the other, plus non-monetary costs*

Broken promises of AIMO

- Vance Packer's *Hidden Persuaders*:
  - *Marketing is good at convincing advertisers and publishers*
  - *Though people are far too smart (they get coffee while the ads plays)*

- Micro-targeting and Behavioural Advertising
  - *Evidence may refer to the selection effect*
  - *Real evidence is absent, if at all attainable*
  - *Some evidence of p-hacking and data dredging*

- So what about AIMO in policing?

DATA FALLACIES TO AVOID

CHERRY PICKING

Selecting results that fit your claim and excluding those that don't.

DATA DREDGING

Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.

SURVIVORSHIP BIAS

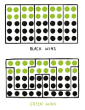
Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.

COBRA EFFECT

Setting an incentive that accidentally produces the opposite result to the one intended. Also known as a Perverse Incentive.

FALSE CAUSALITY

Falsely assuming when two events appear related that one must have caused the other.



GERRYMANDERING

Manipulating the geographical boundaries used to group data in order to change the result.

SAMPLING BIAS

Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.

GAMBLER'S FALLACY

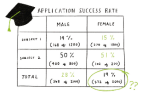
Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).

HAWTHORNE EFFECT

The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.

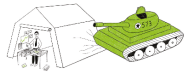
REGRESSION TOWARDS THE MEAN

When something happens that's unusually good or bad, it will revert back towards the average over time.



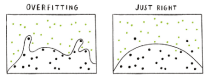
SIMPSON'S PARADOX

When a trend appears in different subsets of data but disappears or reverses when the groups are combined.



MCNAMARA FALLACY

Relying solely on metrics in complex situations and losing sight of the bigger picture.

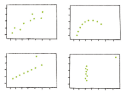


OVERFITTING

Creating a model that's overly tailored to the data you have and not representative of the general trend.

PUBLICATION BIAS

Interesting research findings are more likely to be published, distorting our impression of reality.



DANGER OF SUMMARY METRICS

Only looking at summary metrics and missing big differences in the raw data.

GECKOBOARD.COM

Read more at **data-literacy.geckoboard.com**

Broken promises of AIMO

- Merton, Thomas theorem:
  - *Modulated by Hildebrandt:*
    when machines define a situation as real it is real in its consequences

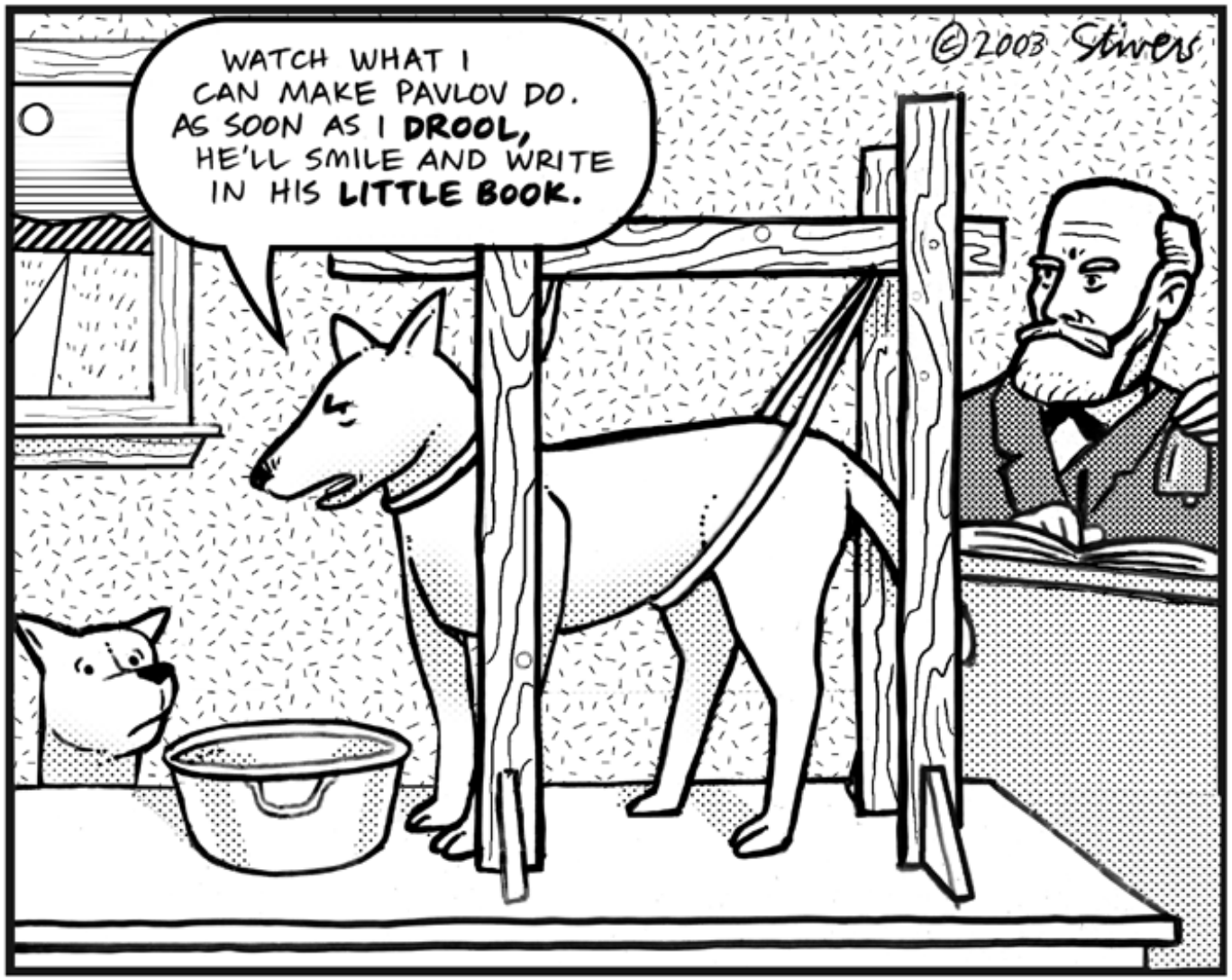
- Goodhart and Campbell effect, Lucas critique:
  - *Articulated by Marylin Strathern*
  - ***When a measure becomes a target, it ceases to be a good measure***

- Esposito's *The Future of Futures* (2011):
  - *Articulated by Hildebrandt:*
    our present futures change the future present

Broken promises of AIMO

- **Underlying reality:**
  - *Double mutual anticipation inherent in human interaction*
  - *Based on birth of the grammatical first person*
  - *G. H. Mead: the I and the me*
    - I am you to you, me to myself
    - You are me to yourself, you are you to me
  - *Parsons and Luhmann: 'double contingency'*

- **Behaviourism** can see the results but cannot grasp the **performative nature of human speech acts**


WATCH WHAT I CAN MAKE PAVLOV DO. AS SOON AS I DROOL, HE'LL SMILE AND WRITE IN HIS LITTLE BOOK.
©2003 Stivers

What's next?

- AIMA and the nature of prediction
- Broken promises?
- AIMA in policing
- Diminishing whose human agency?
- The 'human in the loop'?
- Rule of law and the nature of discretion

AIMA in policing

- **Crime mapping:**
  - *Allocation of resources, sending constables to hotspots*
  - *Based on sensortechnologies and/or crime statistics (smart cities)*

- **Criminal profiling:**
  - *Assessing likelihood of violent conduct or nuisance*
  - *Based on psychometrics and/or crime statistics*

- **Precrime punishment:**
  - *Micro-targeting combined with pre-emption*
  - *Social security and tax fraud detection, stop benefits, impose fines*

AIMA in policing

- Vedder: consequences of non-distributive profiles
  - *Non-monotonic logic, disparate outcomes*
- Harcourt: relative elasticity of policing and offending
  - *Between those profiled and those not profiled*
- Oswald: is use of ML technologies in policing effective?
  - *At group level maybe*
  - *At individual level probably not*

AIMA in policing

- Oswald: bias affects the effectiveness and fairness of smart policing
  - *Not always a matter of 'fixing' the data*
  - *Framing the problem in terms of 'low level nuisance' creates bias*
  - *Funding for technological solutions (while spending cuts for others)*
  - *Opinion-based rather then evidence based (lack of testing)*
  - *Issues around proprietary algorithms*
  - *Amplification of bias due to feedback loops within biased data set*

What is predicted here is future policing, not future crime

What's next?

- AIMA and the nature of prediction
- Broken promises?
- AIMA in policing
- Diminishing whose human agency?
- The 'human in the loop'?
- Rule of law and the nature of discretion

Diminishing whose human agency?

- **Agency** is the ability to navigate the world (action and perception), taking into account the effects of own behaviour
  - *Thermostat, software agent (webbot), self-driving car, plants, animals*

Diminishing whose human agency?

- **Human agency** is the capability to navigate
  - physical space and
  - our shared institutional world,
  - anticipating how we are anticipated

  - *From information to meaning, from brute to institutional facts*

Diminishing whose human agency?

- **Human autonomy** depends on
- the affordances of the environment
- the freedom from unreasonable constraints on

  the construction of our identity (freedom to)
  - *Affordances of an environment are agent dependent*
  - *Capabilities of a human agent are environment dependent*

Diminishing whose human agency?

- Big Data Space
  - *Distributed space*
  - *Distributed storage, distributed access, distributed feedback loops*
  - *A space that de- and re-contextualises data and inferences*

Diminishing whose human agency?

- Digital Unconscious
  - *Ubiquitous, surreptitious, invisible, calm computing*
  - *Nudge theory & ML: pre-emption*
  - *Environment is sticky: it transforms in function of our behaviours*

Diminishing whose human agency?

- Onlife world
  - *Distinction between online and offline becomes artificial*
  - *Environment develops a new mindless type of agency*

What about the conuncrum of the ‘human in the loop’?

Dutch Minister of Justice and Safety on use of AI by the police:

- AI will make a range of pivotal decisions
  - *Mostly in the background*
- The system should be verifiable and results auditable
  - *For decisions with major impact human intervention remains important*

Diminishing whose human agency?

Art. 22 GDPR may provide more protection here than art. 6 ECHR:

- A right not to be subject to:
  - *A decision*
  - *based solely on automated processing, including profiling,*
  - *which produces legal effects concerning him or her*
  - *or similarly significantly affects him or her*

Diminishing whose human agency?

Art. 22 GDPR would provide more protection here than art. 6 ECHR:

- Understood as a prohibition, with 3 exceptions:
  1. *Contract (necessity requirement)*
  2. *EU or MS law (including fundamental rights safeguards)*
  3. *Explicit consent (informed, unambiguous)*

Diminishing whose human agency?

Art. 22 GDPR may provide more protection here than art. 6 ECHR:

- In case of contract or consent:
  - *Right to human intervention*
  - *Right to express point of view*
  - *Right of contestation*

Diminishing whose human agency?

Art. 22 GDPR may provide more protection here than art. 6 ECHR:

- In case of any exception, obligation:
  - *To inform about existence of automated decisions in the sense of 22*
  - *To provide meaningful information about the logic of processing*
  - *To inform about the significance and envisaged consequences*

Diminishing whose human agency?

But GDPR is not applicable, see art. 11 PDPD (less protection)

- Default prohibition
  - *Unless based on MS law*
  - *Which included fundamental rights safeguards*
- More pertinent prohibition if based on use of art. 10 PDPD personal data
- Full prohibition of profiling based on art. 10 PDPD that results in discrimination

Diminishing whose human agency?

The concept of 'the human in the loop':

- For decisions with major impact human intervention remains important
- EDPB:
  - *The human must understand the decision*
  - *The human must be competent to take another decision*
  - *No fake human intervention!*

Why speak of the ‘human in the loop’?

We need to develop ‘a machine in the loop’:

- what matters is human interaction
- speaking of a ‘human in the loop’ ignores our agency
- we don’t want to be ‘in the loop’ but ‘in charge’ or ‘interacting with’

Diminishing whose human agency?

Who is we?

- *Criminal offenders?*
- *Suspects of a criminal offence?*
- *Potential suspects of a criminal offence?*
- *Categories of people who may qualify as potential suspects?*
- *Anyone?*
- *All of us?*

Diminishing whose human agency?

Who is we?

- *Those who develop ML for smart policing?*
- *Those who frame the research design and decisional thresholds?*
- *Those who must work with the software?*
  - Are they in the loop, or is the software in the loop?
  - Can they take decisions, or is their decisional space constrained?

Diminishing whose human agency?

Who is we?

- *Who is actually 'doing' the policing:*
- *those in charge of public management*
- *or those charged with*
  - conducting inquiries,
  - enforcing the law and
  - keeping the peace?

Diminishing whose human agency?

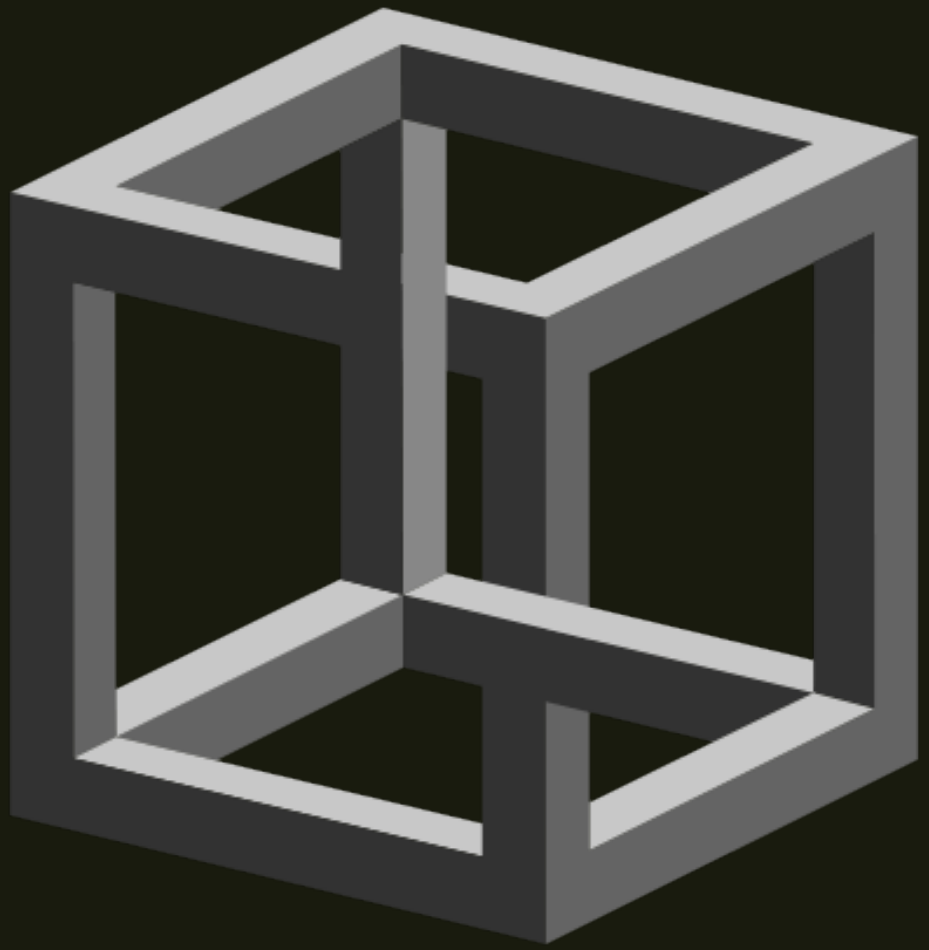
- Street-level policing: exercising discretion, developing acuity
- Screen-level policing: overruled by algorithmic insights
- Lipsky:
  - *Acting based on professional discretion or*
  - *Acting based on imposed rules*
- Evans and Harris:
  - *Discretion is not the absence of principles or rules;*
  - *rather it is the space between them*

What's next?

- AIMA and the nature of prediction
- Broken promises?
- AIMA in policing
- Diminishing whose human agency?
- The 'human in the loop'?
- Rule of law and the nature of discretion

Rule of law and the nature of discretion

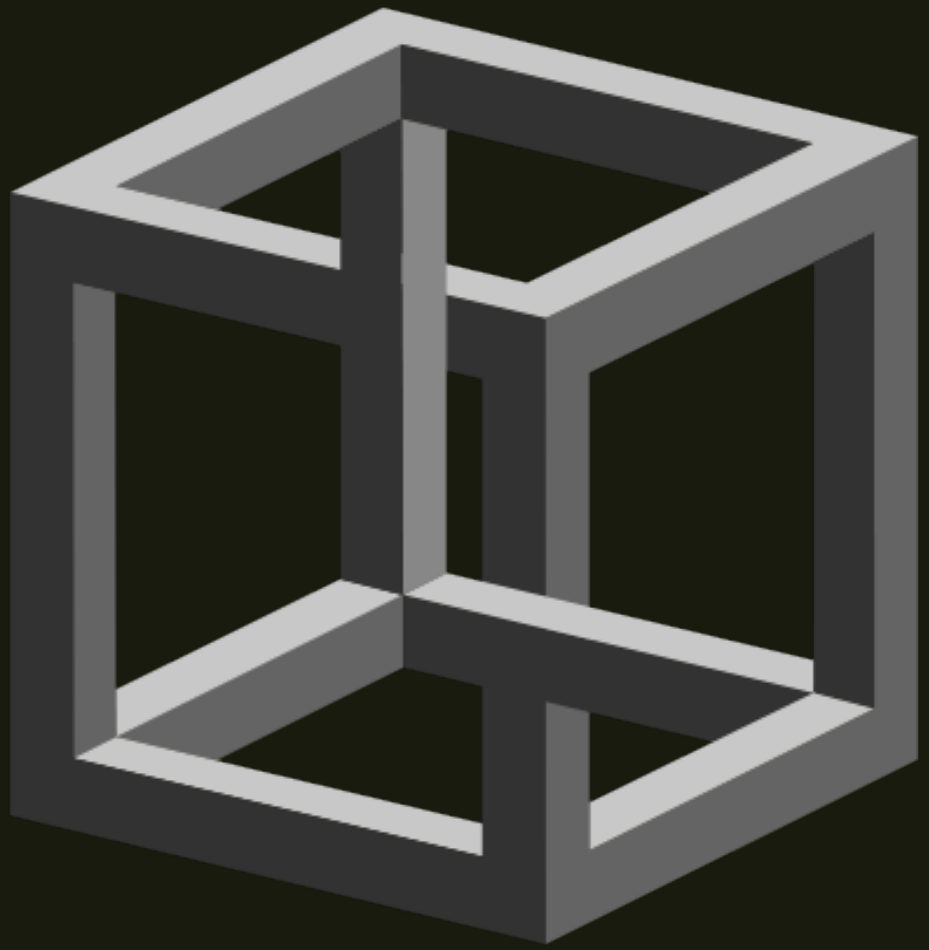
- On discretion (e.g. Dworkin):
  - *Without rules discretion makes no sense*
  - *Discretion combines trans-local rules with local knowledge*
  - *Discretionary decision-making implies interpretation in light of applicable norms*
  - *And interpretation of those norms in light of the situation at hand*
  - *Arbitrary decision-making is the opposite of discretionary decision making*



Two ways of perceiving the same thing:
the rule-perspective
the local perspective

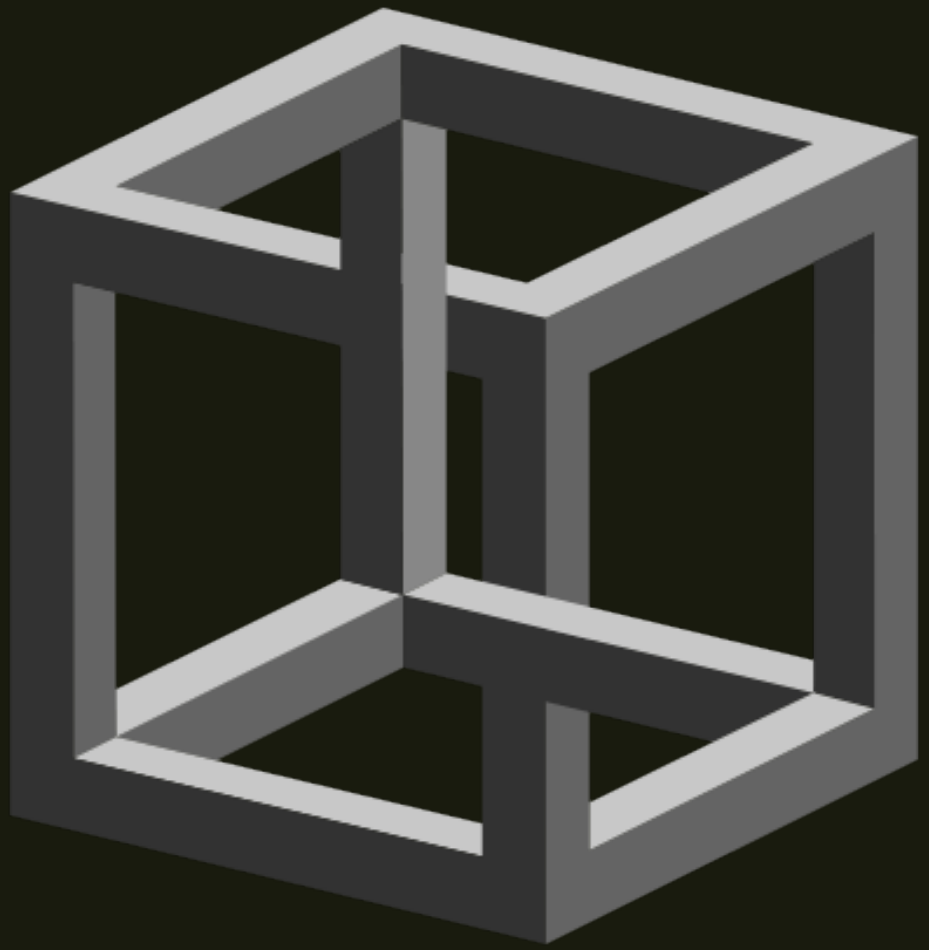
rules make foreseeable
situations are unique

Discretion honours both



Legalism:
mechanical
rule-application

Legality:
safeguarding
legal certainty
justice
instrumentality



We do not want to be the ‘humans in the loop’:

reigned in by

reasonless
uncaring
cybernetic systems

Concluding remarks

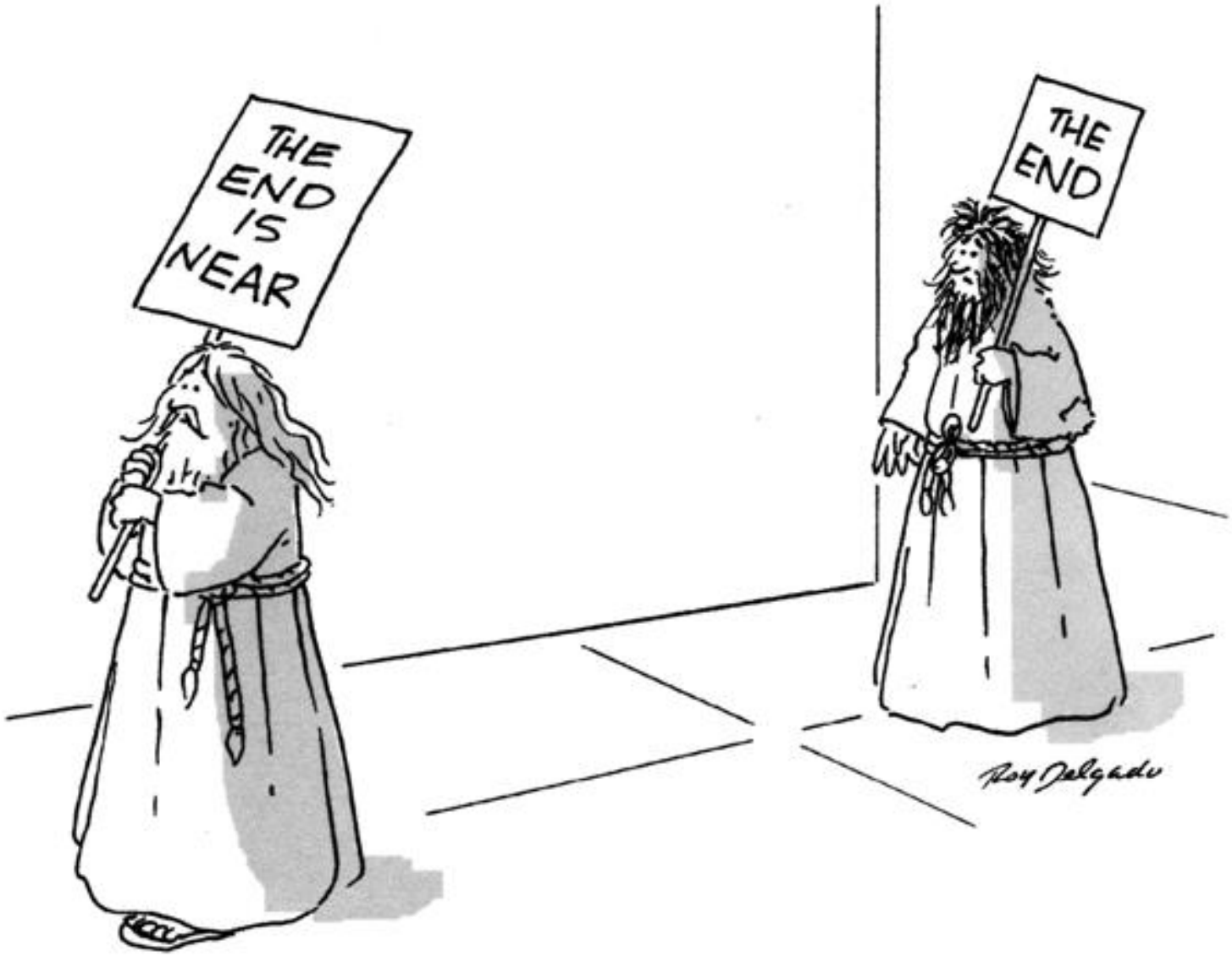
- If data-driven policing operates in terms of 'a human in the loop'
- AI & Big Data in the context of policing may be
  - *Unreliable*
  - *Externalising costs (bias, privacy)*
  - *Incontestable for police and those targeted*
  - *Deskilling police constables*
  - *Disrupt rule of law*

Concluding remarks

- If data-driven policing operates by way of 'a machine in the loop'
- AI & Big Data in the context of policing may be
  - *A way to address bias and prejudice*
  - *A tool to uncover blind spots*
  - *Evidence to uphold the presumption of innocence*
  - *A way to initiate agonistic debate within the police about effectiveness*

Concluding remarks

- If data-driven policing operates by way of 'a machine in the loop'
- AI & Big Data in the context of policing may be beneficial if:
  - *Not used to replace street level constables*
  - *Not employed to deskill officers*
  - *Not used to surveil and control professional police*
  - *Built on agonistic research design*


THE END IS NEAR
THE END
Roy Delgado

- Hildebrandt, M. 2016. 'New Animism in Policing: Re-Animating the Rule of Law?' In *The SAGE Handbook of Global Policing*, edited by Ben Bradford, Beatrice Jauregui, Ian Loader, and Jonny Steinberg, 406–28. S.l.: SAGE Publications Ltd.