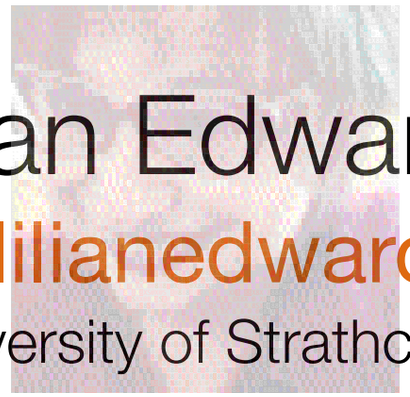
DEPARTMENT OF SCIENCE, TECHNOLOGY,
ENGINEERING AND PUBLIC POLICY

UCL

*Slave to the **algorithm**?*

Why a ‘right to explanation’ is probably not the remedy you’re looking for.



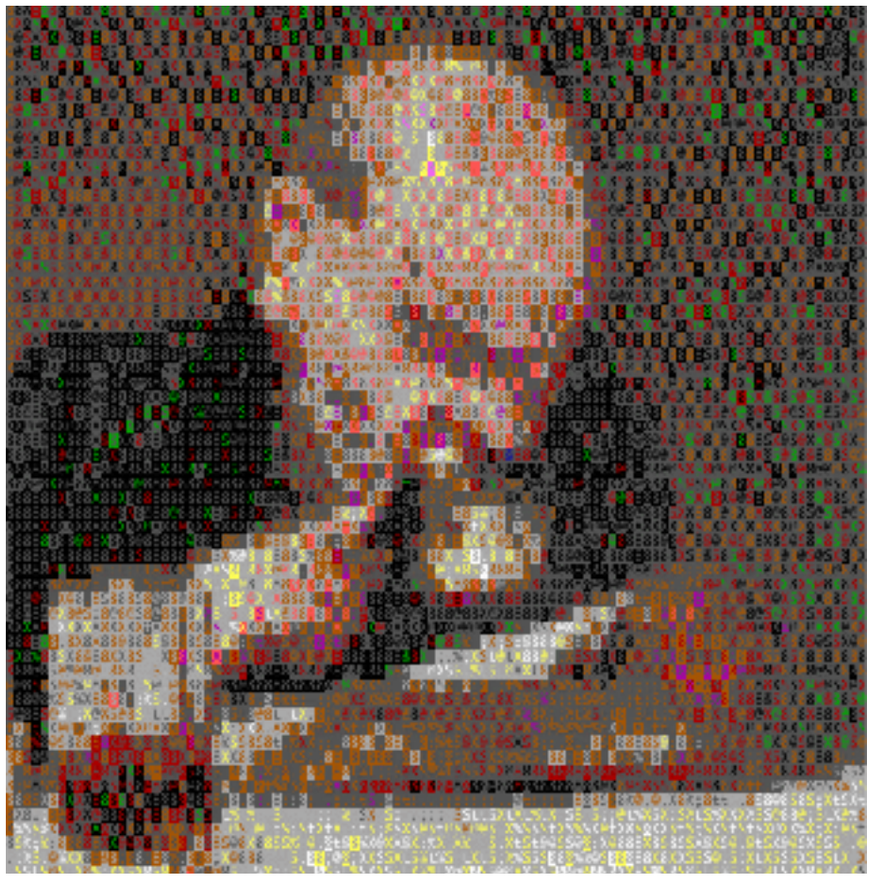
Lilian Edwards
[@lilianedwards]
University of Strathclyde

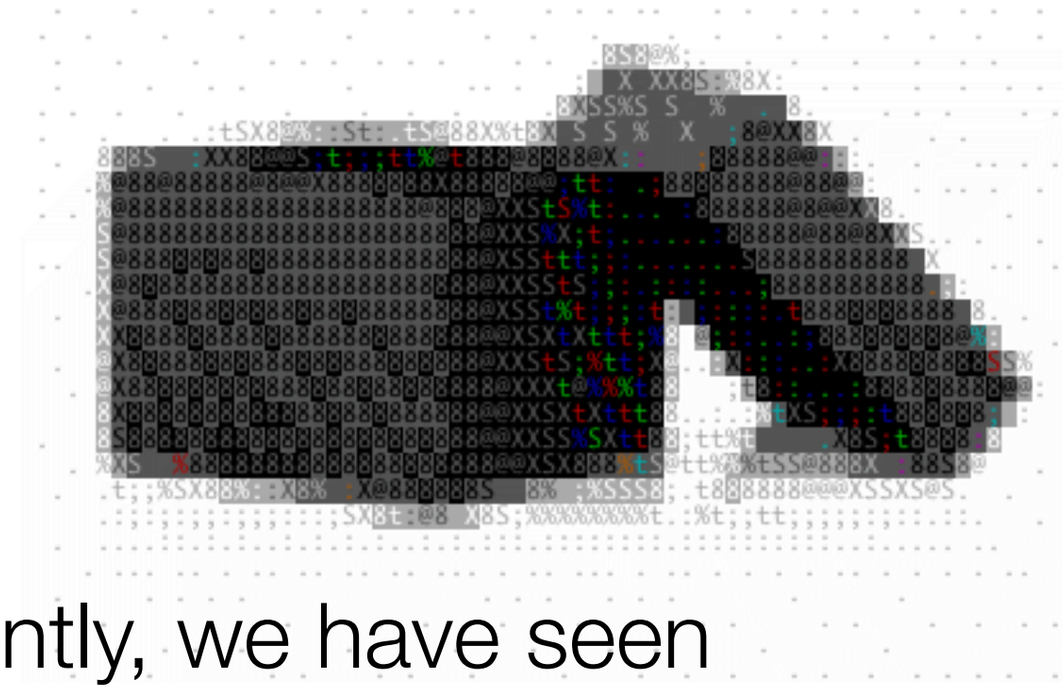
Michael Veale
[@mikarv]
University College London

Big Data: New challenges for law and ethics
Faculty of Law, University of Ljubljana, 22 May 2017

Overview

Concerns around “**black box**”, algorithms, increasingly machine learning algorithms that improve with data, **have fuelled calls to “open them up”.**





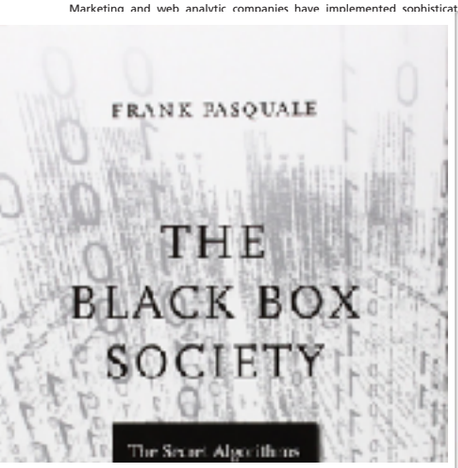
Consequently, we have seen manhunt in data-related laws: **are there provisions that give us a practical right to have these systems “explained”?**

algorithms, everywhere

A New Algorithmic Identity
Soft Biopolitics and the Modulation of Control

John Cheney-Lippold

Abstract
Marketing and web analytic companies have implemented sophisticated



Original Research Article

Heuristics of the algorithm: Big Data, user interpretation and institutional translation

Göran Bolin and Jonas Andersson Schwarz

Abstract
Intelligence on mass media audiences was founded on representative statistical sam
market departments of media corporations. The techniques for aggregating user
ubiquitous personal media (e.g. laptops, smartphones, credit cards/swipe cards

Synthese (2009) 169:593–613
DOI 10.1007/s11229-008-9438-z

The philosophy of simulation: hot new issues or same old stew?

Roman Frigg · Julian Reiss

Received: 5 December 2007 / Accepted: 17 October 2008 / Published online: 4 December 2008

Abstract Computer simulations are an exciting tool that plays important roles in many scientific disciplines. This has attracted the attention of a number of philosophers of science. The main tenor in this literature is that computer simulations not only constitute interesting and powerful new *science*, but that they also raise a host of new *philosophical* issues. The protagonists in this debate claim no less than that simulations call into question our philosophical understanding of scientific ontology, the epistemology and semantics of models and theories, and the relation between experimentation and theorising, and submit that simulations demand a fundamentally new philosophy of science in many respects. The aim of this paper is to critically evaluate

The Relevance of Algorithms

Tarleton Gillespie

forthcoming, in *Media Technologies*, ed. Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot. Cambridge, MA: MIT Press.

Algorithms play an increasingly important role in selecting what information is considered most relevant to us, a crucial feature of our participation in public life. Search engines help us navigate massive databases of information, or the entire web. Recommendation algorithms map our

Equality of Opportunity in Supervised Learning

Moritz Hardt Eric Price Nathan Srebro

October 11, 2016

Abstract

We propose a criterion for discrimination against a specified sensitive attribute in supervised learning, where the goal is to predict some target based on available features. Assuming data about the predictor, target, and membership in the protected group are available, we show how to optimally *adjust* any learned predictor so as to remove discrimination according to our definition. Our framework also improves incentives by shifting the cost of

INTERNET POLICY REVIEW
Journal on internet regulation
Volume 2 | Issue 3

Governance by algorithms

Francesca Musiani
MINES ParisTech , France , francesca.musiani@mines-paristech.fr

Published on 09 Aug 2013 | DOI: 10.14763/2013.3.188

Abstract: Algorithms are increasingly often cited as one of the fundamental shaping devices of our daily, immersed-in-information existence. Their importance is acknowledged, their performance scrutinised in numerous contexts. Yet, a lot of what constitutes 'algorithms' beyond their broad definition as "encoded procedures for transforming input data into a desired output, based on specified calculations" (Gillespie, 2013) is often taken for granted. This article seeks to

Article

Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness

Science, Technology, & Human Values
1-25

DOI: 10.1177/0162243915606523
sthv.sagepub.com
SAGE

Mike Ananny[1]

Abstract
Part of understanding the meaning and power of algorithms means asking what new demands they might make of ethical frameworks, and how they

Article

Algorithms, Governance, and Governmentality: On Governing Academic Writing

Science, Technology, & Human Values
1-33

DOI: 10.1177/0162243915587360
sthv.sagepub.com
SAGE

Lucas D. Introna[1]

Abstract
Algorithms, or rather algorithmic actions, are seen as problematic because they are inscrutable, automatic, and subsumed in the flow of daily practices. Yet, they are also seen to be playing an important role in organizing opportunities, enacting certain categories, and doing what David Lyon calls "social sorting." Thus, there is a general concern that this increasingly prevalent mode of ordering and organizing should be governed more

Original Article

Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics

Science, Technology, & Human Values
1-16

DOI: 10.1177/0162243915589635
sthv.sagepub.com
SAGE

Kate Crawford[1,2,3]

Abstract
This paper explores how political theory may help us map algorithmic logics against different visions of the political. Drawing on Chantal Mouffe's theories of agonistic pluralism, this paper depicts algorithms in public life in ten distinct scenes, in order to ask the question, what kinds of politics do they instantiate? Algorithms are working within highly contested online spaces of public discourse, such as YouTube and Facebook, where incompatible perspectives coexist. Yet algorithms are designed to produce clear "win-

Article

Want to be on the top? Algorithmic power and the threat of invisibility on Facebook

Taina Bucher
University of Oslo, Norway

Abstract
This article explores the new modalities of
a focus on the social networking site Face
Panopticism – that is, the architectural st

Governing Algorithms: A Provocatio

Solon Barocas Sophie Hood Malt
sib237@nyu.edu sah7@nyu.edu mcz3

New York University
March 29, 2013

Paper prepared for the "Governing Algorithms" conference, May 16-17, 2013, at New York University

1. Algorithms have developed into somewhat of a modern myth. They "compet
rooms" (Slavin 2011), "determine how a billion plus people get where they're g
2011), "have already written symphonies as moving as those composed by Beet
2012), "allow self-determined action on the Internet but also contain aspects of
action" (Institute for Media Archeology 2011), and "free us from sorting throug
irrelevant results" (Spring 2011). How do they do all this, and more? What exa

Ethics Inf Technol (2011) 13:251–260
DOI 10.1007/s10676-010-9233-7

Is there an ethics of algorithms?

Felicitas Kraemer · Kees van Overveld · Martin Peterson

Published online: 3 July 2010

Abstract We argue that some algorithms are value-laden, and that two or more persons who accept different value-judgments may have a rational reason to design such algorithms differently. We exemplify our claim by dis-

detail how to solve a problem. Both c
use algorithms for solving a wide
However, in this paper we shall be c
with algorithms implemented in com

ALGORITHMIC ACCOUNTABILITY
Journalistic investigation of computational power structures

Nicholas Diakopoulos

Every day automated algorithms make decisions that can amplify the power of businesses and

BAE Systems' Taranis drone has autonomous elements, but relies on humans for combat decisions.

Ethics of artificial intelligence

Four leading researchers share their concerns and solutions for reducing societal risks from intelligent machines.

▶

Article

Bearing Account-able Witness to the Ethical Algorithmic System

Science, Technology, & Human Values
1-27

DOI: 10.1177/0162243915598056
sthv.sagepub.com
SAGE

Daniel Neyland[1]

Abstract
This paper explores how accountability might make ot
and inaccessible algorithms available for governance. The

Certifying and removing disparate impact

Michael Feldman, *Haverford College*
Sorelle A. Friedler, *Haverford College*
John Moeller, *University of Utah*
Carlos Scheidegger, *University of Arizona*
*Suresh Venkatasubramanian**, *University of Utah*

Abstract

What does it mean for an algorithm to be biased? In U.S. law, unintentional bias is encoded via *disparate impact*, which occurs when a selection process has widely different outcomes for different

Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining

Sara Hajian
Eurecat
Barcelona, Spain
sara.hajian@eurecat.org

Francesco Bonchi
ISI Foundation
Turin, Italy
francesco.bonchi@isi.it

Carlos Castillo
Eurecat
Barcelona, Spain
chato@acm.org

ABSTRACT

lgorithms and decision making based on Big Data have be-
me pervasive in all aspects of our daily lives lives (offline
d online), as they have become essential tools in personal
nance, health care, hiring, housing, education, and poli-
es. It is therefore of societal and ethical importance to ask
hether these algorithms can be discriminative on grounds
ch as gender, ethnicity, or health status. It turns out that
e answer is positive: for instance, recent studies in the con-
xt of online advertising show that ads for high-income jobs
e presented to men much more often than to women [5];
d ads for arrest records are significantly more likely to
ow up on searches for distinctively black names [16].
This *algorithmic bias* exists even when there is no discrim-
nation intention in the developer of the algorithm. Some-
mes it may be inherent to the data sources used (soft-
are making decisions based on data can reflect, or even
mplify, the results of historical discrimination), but even
hen the sensitive attributes have been suppressed from the
put, a well trained machine learning algorithm may still
iscriminate on the basis of such sensitive attributes because
f correlations existing in the data. These considerations
ll for the development of data mining systems which are
iscrimination-conscious by-design. This is a novel and cha-
lging research area for the data mining community.

1. INTRODUCTION

At the beginning of 2014, as an answer to the growing concerns about the role played by data mining algorithms in decision-making, USA President Obama called for a review of big data collecting and analysing practices. The resulting report[1] concluded that "big data technologies can cause societal harms beyond damages to privacy." In particular, it expressed concerns about the possibility that decisions informed by big data could have discriminatory effects, even in the absence of discriminatory intent, further imposing less favorable treatment to already disadvantaged groups.

In the data mining community, the effort to design discrimination-conscious methods has developed two groups of solutions: (1) techniques for *discrimination discovery from databases* [13] and (2) discrimination prevention by means of *fairness-aware data mining*, developing data mining systems which are discrimination-conscious by-design [8]. Discrimination discovery in databases consists in the actual discovery of discriminatory situations and practices hidden in a large amount of historical decision records. Discrimination prevention in data mining consists of ensuring that data mining models automatically extracted from a data set are such that they do not lead to discriminatory decisions even if the data set is inherently biased against protected groups. Different discrimination prevention meth-

BIG DATA'S DISPARATE IMPACT

*Solon Barocas**
Andrew D. Selbst†

Big data claims to be neutral. It isn't.

Advocates of algorithmic techniques like data mining argue that they eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with. Data mining can inherit the prejudices of prior decision-makers or reflect the widespread biases that persist in society at large. Often, the "patterns" it discovers are simply preexisting societal patterns of inequality and exclusion. Unthinking reliance on data mining can deny members of vulnerable groups full participation in society. Worse still, because the resulting discrimination is almost always an unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers, it can be unusually hard to identify the source of the problem or to explain it to a court.

This Article examines these concerns through the lens of American anti-discrimination law—more particularly, through Title VII's prohibition on discrimination in employment. In the absence of a demonstrable intent to discriminate, the best doctrinal hope for data mining's victims would seem to lie in disparate impact doctrine. Case law and the EEOC's Uniform Guidelines, though, hold that a practice can be justified as a business necessity where its outcomes are predictive of future employment outcomes, and data mining is specifically designed to find such statistical correlations.

Mireille Hildebrandt • Serge Gutwirth
Editors

Profiling the European Citizen

Cross-Disciplinary Perspectives

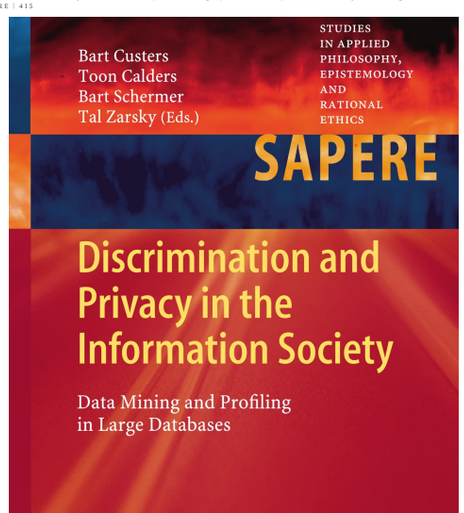
COMPUTING AND ACCOUNTABILITY

Helen Nissenbaum

A teacher stands before her sixth-grade class demanding to know who shot the spitball in her ear. She threatens punishment for the whole class if someone does not step forward. Eyes are cast downward and nervous giggles are suppressed as a boy in the back row slowly raises his hand.

The boy in the back row has answered for his actions. We do not know whether he shot at the teacher intentionally or merely missed his true target, whether he acted alone or under goading from classmates, or even whether the spitball was in protest for an unreasonable action taken by the teacher. While all of these factors are relevant to determining a just response to the boy's action, the boy, in accepting responsibility for his action, has fulfilled the valuable social obligation of accountability.

In an increasingly computerized society, where computing, and its broad application, brings dramatic changes to our way of life, and exposes us to harms and risks, accountability is very important. A community (a society or professional community) that insists on accountability, in which agents are expected to answer for their work, signals esteem for high-quality work, and encourages diligent, responsible practices. Furthermore, where lines of accountability are maintained, they provide the foundations for just punishment as well as compensation for victims. By contrast, the absence of accountability means that no one answers for harms and risks. Insofar as they are regretted, they are seen as unfortunate accidents—consequences of a brave new technology. As with



What's all this about a "right to explanation"?

- A widely circulated ML confererence paper* superficially read the **General Data Protection Regulation** (GDPR) to claim the existence of new "right to explanation" that would challenge black-boxed machine learning in practice.
- This stream of argument had been published several times before, regarding a very similar provision in the Data Protection Directive 1995.**
- *In this paper we argue*:
  - Explanations in the GDPR are limited in many important respects.
  - Even overcoming these limitations, the provisions in the GDPR fit poorly with realistic possibilities for explanation facilities from computer science.
  - Limitations to realistic explanation facilities limit the usefulness of this remedy
  - Rights-based explanation remedies fail to address the more common group-related harms we see in practice.

GDPR, art 22

Automated individual decision-making, including profiling

The GDPR ported, with little change, a corresponding right from the 1995 Data Protection Directive (DPD) art 15 to a new art 22 which provides

> *"the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects, concerning him or her, or significantly affects him or her"*

First-glance limitations of GDPR, art 22

Automated individual decision-making, including profiling

Scholars have argued* that article 22 (formerly article 15, DPD), is a "*second class data protection right… rarely enforced and easily circumvented*". What are its characteristics?

1. Not an explanation remedy: limited to:
    i. **preventing certain processing**;
    ii. inserting a **human-in-the-loop**;
    iii. *implicitly,* a **right to be informed** when an automated decision is being made.

Furthermore, only applies to

2. decisions made "**solely on automated processing**"
3. decisions which produce **"legal effects […] or significantly affects him or her"**

Zooming in: ‘Decisions’ in the GDPR

- Is the output of an ML model always a “decision”? GDPR silent on what a decision is, other than it “may include a measure”.

- In credit scoring, as the foundational paradigm, there is no doubt that there is a decision (by the credit offering company) and that it affects an individual data subject (the person seeking credit).

- But in the many well-known “war stories” of algorithmic harms that have been highlighted by scholars and the media, this is not so clear. In these cases, **harms are often not against an individual, but against a group**.

Algorithmic ‘war-story’ I

Professor Latanya Sweeney and Google AdSense

- It’s 2013, and Sweeney, a researcher at Harvard University, investigated the delivery of targeted adverts by Google AdSense using a sample of racially associated names.

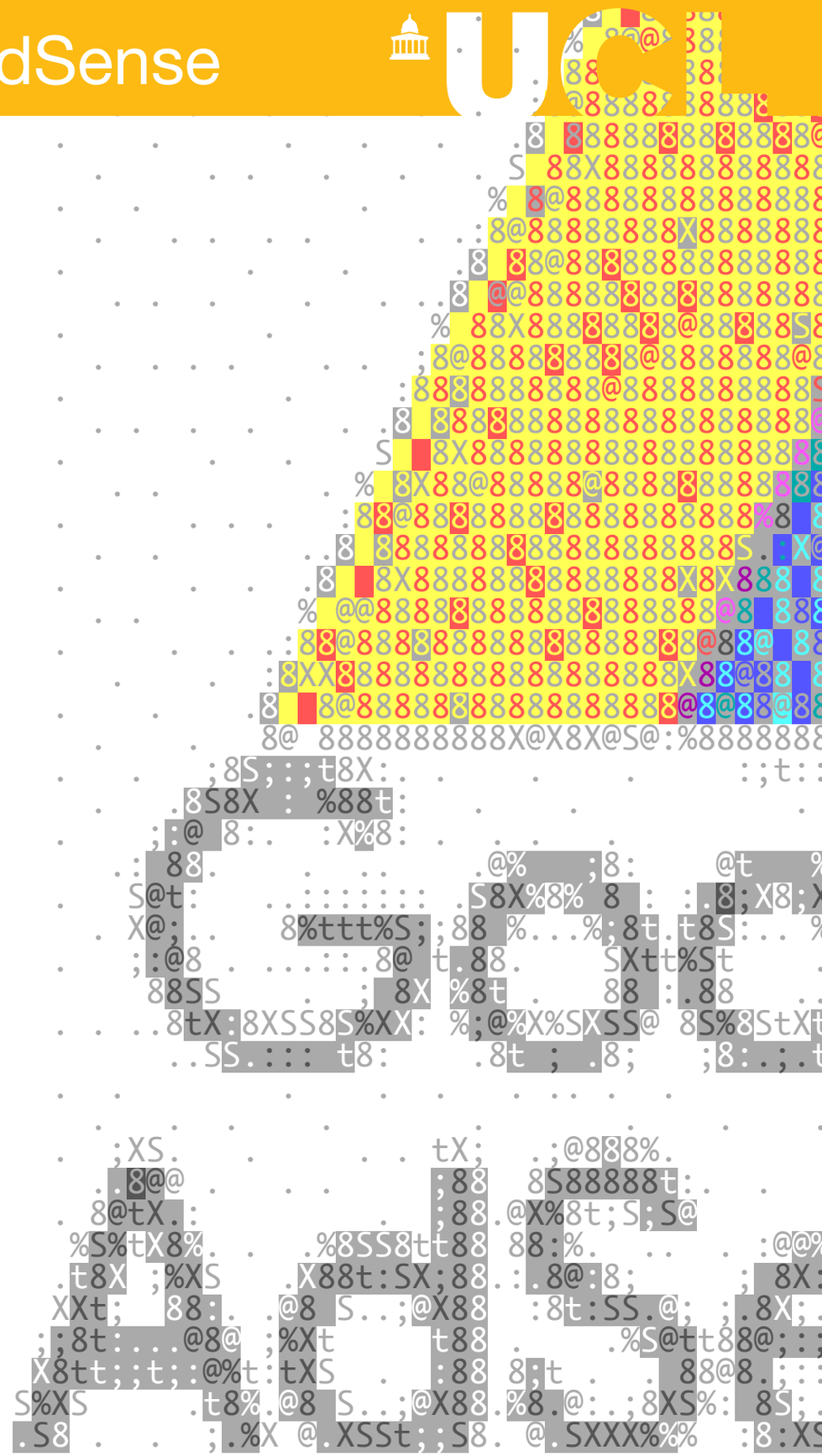
- First names associated predictively with non-white racial origin generated a far higher percentage of adverts associated with or using the word “arrest” when compared to ads delivered to “white” first names.

Algorithmic ‘war-story’ I

Professor Latanya Sweeney and Google AdSense

Was a “decision” taken with reference to Sweeney?

- No effect on legal status (public status, such as citizenship, or private status, like capacity to make a will)
- A **so-called racial group was impugned by assumption of above average criminality**: takes us to a sort of ‘group right’, very different from individual liberal paradigm rights granted by the GDPR.
- Even given impact on Sweeney as an individual constructed through group membership, was it “significant”?



Algorithmic ‘war-story’ II
“Jew Watch”

- In 2004, the Google search algorithm(s) placed a site “Jew Watch” at the top of the rankings for many people who searched on the word “Jew”. Google pulled what we now might call the “neutrality” defence, or fallacy.

- Was there a significant decision made? Only individuals searching “Jew” could be offended. While in algorithmic defamation cases, searched names are clearly linked to individuals, here, the searched term is linked broadly to a group.

Back to GDPR, art 22

Automated individual decision-making, including profiling

1. The data subject shall have the 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.

2. Paragraph 1 shall not apply if the decision:

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

(c) is based on the data subject's explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

The pardoxes of GDPR, art 22

Automated individual decision-making, including profiling

When sensitive data, such as so-called race, is being processed, you do not have a right to have a human-in-the-loop if

- Explicit consent was given (or substantial public interest in basis of MS law);
- Safeguards in Recital 71 are in place, which include

> the controller, or necessary for the entering of performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent. In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision. Such measure should not concern a child.

Under art 22, it seems possible to infer that **where you have no primary right to object, you have a binding right to explanation.** Else, no. Paradoxical, in ways. To trigger such a right, you need to know that sensitive data — in the war stories, inferred sensitive data — were being processed. How, without a right to explanation?

Additionally, challenge

Very shaky grounds to found EU transparency rights upon.*

* For more, including an analysis of case law, see Wachter et al. (forthcoming) Why a right to explanation does not exist in the General Data Protection Regulation, *International Data Privacy Law*

Continuing our quest, with GDPR, art 15?

Right of access by the data subject

Article 15 — which existed in the DPD as article 12(a) — provides that the data subject shall have the right to confirm whether or not personal data relating to him or her are being processed by a controller and if that is the case, access to that personal data and the "following information."

This includes in the context of "automated decision making […] referred to in art 22(1) and (4)" access to "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing" (art 15(1)(h)).

It has its own problems in the form of a carve out for the protection of trade secrets and intellectual property, although Recital 63 now counsels that this should not justify "a refusal to provide *all* information to the data subject" (emph added)

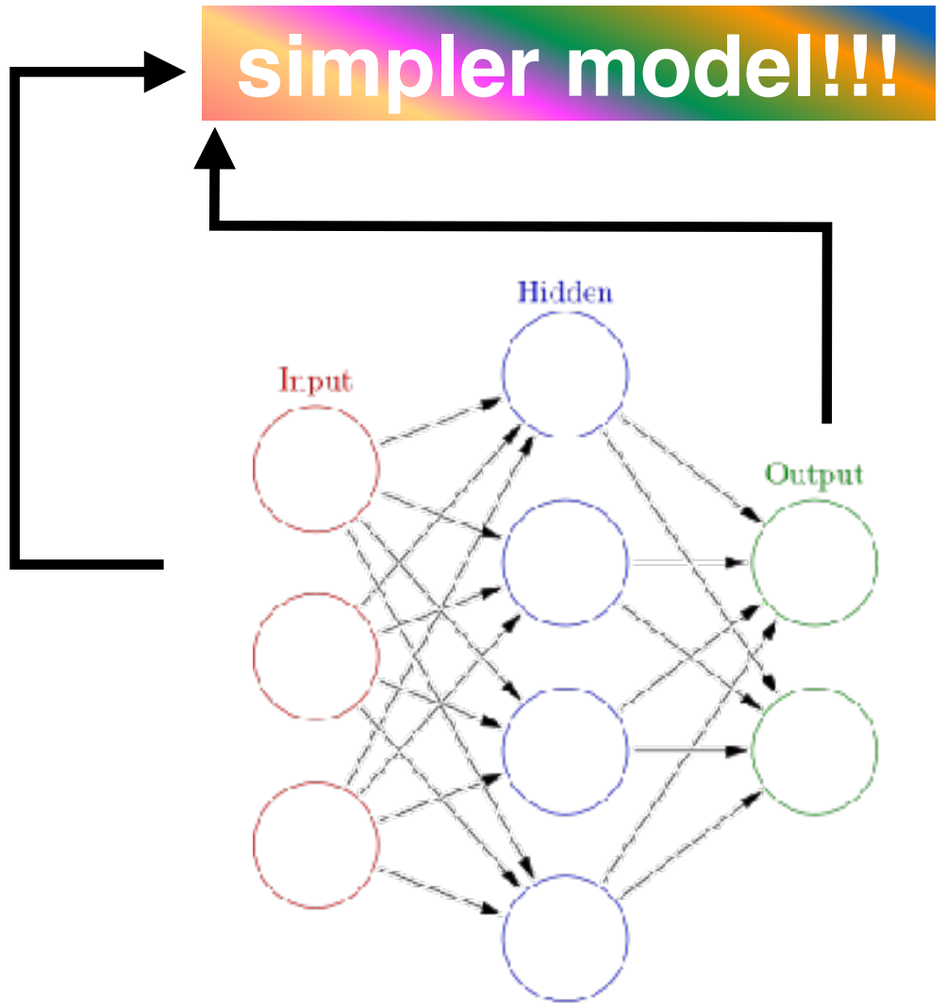
How far must we breach IP for our beloved explanations?

A trip to old computer science literature

- Work on explaining machine learning in the early 90s, in the context of “expert systems”, concluded that a “trace” of what the computer did was not useful. **Explanation was a separate optimisation task.***

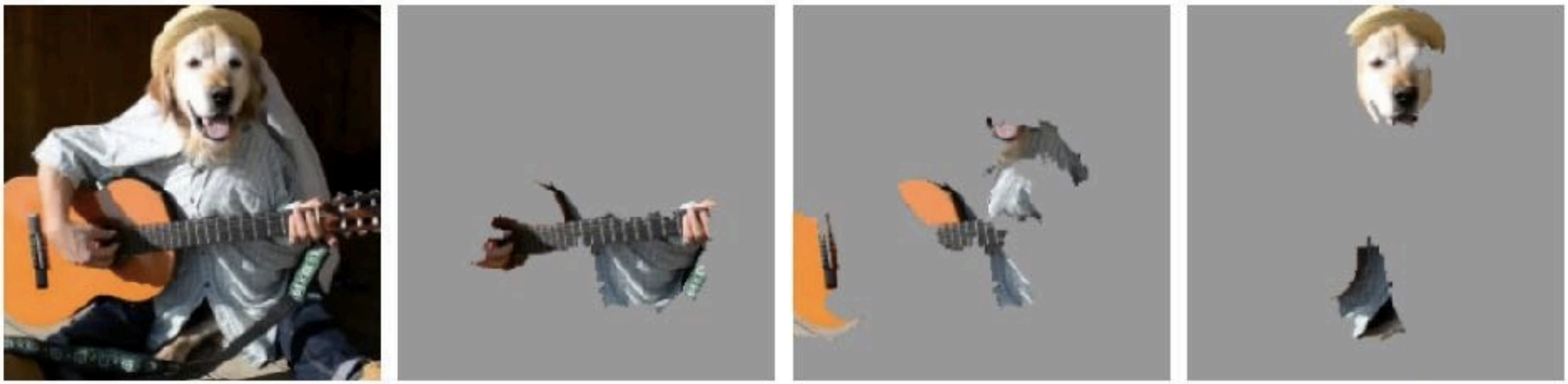
Two options:

- could reach in (decompose) the complex ML system
- could fit a simpler model around the black box to estimate its core logics (pedagogical approach).**



*Wick and Thompson (1992) (doi:10.1016/0004-3702(92)90087-E); **Tickle et al. (1998) (doi:10.1109/72.728352)

And yes, this works today*



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

*Ribeiro et al. (2016) (doi:10.1145/2939672.2939778). Decompositional methods might overtake the effectiveness of this. See Montavon et al. (2017) (doi:10.1145/2939672.2939778)

An important distinction

Model-centric explanations

"Global", attempts to be equally valid for a whole model

- setup information
- training metadata
- performance metrics
- estimated global logics

Subject-centric explanations

"Local" around a particular question or input

- sensitivity-based
- case-based
- demographic-based
- performance-based

Great news for IP and trade secrets, right?!

- Yes — the IP/trade secrets defence is weakened from both practical and legal angles.

- Yet, we face other practical challenges that might supper our success

What might scupper our rights to explanation?

Domain

Some tasks are easier to ‘explain’ than others. When you’re using only a few input variables, not too tricky.

When you’re using a lot — say, thousands

- it’s quite easy if we can mentally compress them, like pixels in an image

- but much harder when they’re quite abstract, like many sensor readings, or clicks/browser history online, or GPS movements, as they just overwhelm us.

What might scupper our rights to explanation?

Users

As discussed, pedagogical approaches make a more “simple” model, that is more interpretable. What do they simplify?

Who is more likely to call upon a right of explanation? Probably, someone who has been treated weirdly by a model. This weirdness is probably complex, rather than simple.

A RtE might fail on those who call on it most.

Using other GDPR rights

Right to erasure; Right to data portability

GDPR, art 17
Right to erasure

- ML as dataset: ML can ‘leak’ personal data*. Want the model amended so it cannot.
- ML contains undesirable inferences about me. Want these removed from a model.

Yet!

- Impracticality of *machine unlearning*
- Collective action problem
- Erasure from traded models

GDPR, art 20
Right to data portability

- Take your inferences — the value from your data — to more ethical decision-makers.
- Download these inferences in a machine readable form — an explanation?

Yet!

- Article 29 Working Party**: *provided* includes *observed*, but not *inferred*
- Consumer inertia

The GDPR mandates **data protection impact assessments** in particular where there is

"systematic and extensive evaluation of personal aspects relating to natural persons […] based on automated processing, including profiling […] and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person." [GDPR, art 35(3)(a)]

Where "high risk", the local member state DPA must be consulted before the system can be launched. The impact assessment must be shared and the DPA must provide written advice to the controller and can use their powers to temporarily or permanently ban use of the system.

Making better algorithms with GDPR tools
Privacy seals

The GDPR supports voluntary certification arrangements, to be developed by DPAs.

This could be used to create more sector-specific outlines for ML systems in different domains.

Also provide access requirements to data and logics for bodies representing consumers, rather than individuals.

Similarities to the “supercomplaint” systems in UK finance and consumer law.

thanks — q?

Draft paper

`http://michae.lv/explanations.pdf`

Secret Ljubljana link!

Will be on SSRN in a fortnight after workshopping

lilian edwards // @lilianedwards
lilian.edwards@strath.ac.uk

michael veale // @mikarv
m.veale@ucl.ac.uk

EPSRC
Engineering and Physical Sciences Research Council