Overview

▪ Fairness and AI
▪ Explanations and AI
▪ Transparency and AI
▪ Solutions?
I. Automated decision-making and AI

- Workplace
- Judicial system, criminal justice
- Health sector
- Transport
- Insurance
- Finance (trading)
- Banking (loans)
- Crime
Advantages

- Accuracy
- Spot new patterns
- Time and cost efficient
- Consistent
AI and the Black Box

- Opaque
- Difficult to explain
- Inscrutable
- Bias, discrimination
What do we expect of AI?

- Fair
- Explainable
- Transparent
Fairness and AI
That was surprisingly easy. How come the robotic uprising used spears and rocks instead of missiles and lasers?

If you look to historical data, the vast majority of battle-winners used pre-modern weaponry.

Thanks to machine-learning algorithms, the robot apocalypse was short-lived.
Inferential analytics

- Assessments (e.g. settlements)
- Predictions (e.g. trial outcomes, risk scores)
AI predicts outcomes of human rights trials

24 October 2016

The judicial decisions of the European Court of Human Rights (ECtHR) have been predicted to 79% accuracy using an artificial intelligence (AI) method developed by researchers at UCL, the University of Sheffield and the University of Pennsylvania.
Classification: Wolf or a Husky?

Only 1 mistake!
Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusted the bad model</td>
<td>10 out of 27</td>
<td>3 out of 27</td>
</tr>
<tr>
<td>Snow as a potential feature</td>
<td>12 out of 27</td>
<td>25 out of 27</td>
</tr>
</tbody>
</table>

Table 2: “Husky vs Wolf” experiment results

learning, specifically for vision tasks [3] know when the systems are likely to increase in trust, by avoiding “silly” solutions either require additional annoying engineering that is specific to vision tasks. Insight into why a decision should not more, they assume that the current evaluation is reliable, which may not be the case if privacy leakage are present. Other recent works propose different kinds of mistakes. Interestingly, the subjects in their studies showed serious problems in the 20 newsgroups data test sets in these tasks, suggesting that increased learning is not sufficient. Note that Groce et al. in this regard, many researchers in the field have published classifiers that would not generalise using LIME, we show that even non-identifiable these irregularities when explaining. Further, LIME can complement these existing methods to allow users to assess trust even when “accurate” but incomplete, for other applications.
Racial bias in police stop and search getting worse, report reveals

Despite reforms, black people are nine times more likely than white people to be checked for drugs
Machine Bias

There's software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016
Explanations and AI
YOU WOULDN'T UNDERSTAND!

How can you judge me?
Explaining Explanations in AI


10 Pages  •  Posted: 27 Nov 2018

Brent Mittelstadt
University of Oxford - Oxford Internet Institute

Chris Russell
University of Surrey

Sandra Wachter
University of Oxford - Oxford Internet Institute

Date Written: November 4, 2018

Abstract

Recent work on interpretability in machine learning and AI has focused on the building of simplified models
What is an explanation in ML?

Mostly scientific models: decision trees; (linear) approximations for debugging

“All models are wrong but some are useful”
(George E. P. Box)
What is an explanation in philosophy & cognitive science?

“An explanation ought to be such that it enables us to see what sort of difference it would have made for the outcome if the factors cited in the explanation had been different in various possible ways”

(Woodward 2003)

‘Everyday explanations’ are contrastive, selective, and social.’

(Miller 2017)
Challenges for explanations

▪ Might infringe trade secrets or privacy
▪ Technically unfeasible
▪ Might not be meaningful for individuals
▪ Manipulation and gaming of the system
Harvard Journal of Law & Technology
Volume 31, Number 2 Spring 2018

COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR

Sandra Wachter,* Brent Mittelstadt,** & Chris Russell***

TABLE OF CONTENTS

I. INTRODUCTION ............................................................................................................. 842

II. COUNTERFACTUALS .................................................................................................... 844
   A. Historic Context and the Problem of Knowledge ................................................... 846
   B. Explanations in A.I. and Machine Learning ......................................................... 849
   C. Adversarial Perturbations and Counterfactual Explanations ............................... 851
   D. Causality and Fairness ............................................................................................ 853
Three goals of explanations

"You were denied parole because you had 3 prior offences. If you have had 2 prior offences, you would have been granted parole."

- Understand decisions
- Challenge decisions
- Alter future decisions
What If...

you could inspect a machine learning model, with no coding required?

Building effective machine learning systems means asking a lot of questions. It's not enough to train a model and walk away. Instead, good practitioners act as detectives, probing to
Transparency and AI
Trade secret cases

- Loomis v. Wisconsin, 2017
- Medicaid program, Idaho, 2016
- Houston Federation of Teachers vs. Houston Independent School District, 2017
I

(Legislative acts)

DIRECTIVES

DIRECTIVE (EU) 2016/943 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL
of 8 June 2016
on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure
(Text with EEA relevance)
Trade Secret Directive - What is a trade secret?

a) it is secret in the sense that it is not, [...] known among or readily accessible to persons within the circles that normally deal with the kind of information in question;

(b) it has **commercial value** because it is secret;

(c) it has been subject to **reasonable steps** under the circumstances, by the person lawfully in control of the information, **to keep it secret**;
Trade secrets

Among other things:

- Data
- Assessments and predictions (risk assessments)
- Code
- Algorithms
- Model
28th of September 2011

From: Facebook
To: Max Schrems

Dear Mr. Schrems:

We refer to our previous correspondence and in particular your subject access request dated July 11, 2011 (the Request).

To date, we have **disclosed all personal data to which you are entitled** pursuant to Section 4 of the Irish Data Protection Acts 1988 and 2003 (the Acts).

Please note that certain categories of personal data are exempted from subject access requests. Pursuant to Section 4(9) of the Acts, personal data which is **impossible to furnish or which can only be furnished after disproportionate effort** is exempt from the scope of a subject access request. We have not furnished personal data which cannot be extracted from our platform in the absence of disproportionate effort.

Section 4(12) of the Acts carves out an exception to subject access requests where the disclosures in response would adversely affect **trade secrets or intellectual property**. We have not provided any information to you which is a trade secret or intellectual property of Facebook Ireland Limited or its licensors.

Please be aware that we have complied with your subject access request, and that we are not required to comply with any future similar requests, unless, in our opinion, a reasonable period of time has elapsed.

Thanks for contacting Facebook,
Facebook User Operations Data Access Request Team
Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation

Sandra Wachter*, Brent Mittelstadt** and Luciano Floridi***

Key Points

- Since approval of the European Union General Data Protection Regulation (GDPR) in 2016, it has been widely and repeatedly claimed that a ‘right to explanation’ of all decisions made by automated or artificially intelligent algorithmic systems will be legally mandated by the GDPR once it is in force, in 2018.

- The ambiguity and limited scope of the ‘right not to be subject to automated decision-making’ contained in Article 22 (from which the alleged ‘right to explanation’ stems) raises questions over the protection actually afforded to data subjects.

- These problems show that the GDPR lacks precise
Art 12a Data Protection Directive

Right of access

right to obtain from the controller - “knowledge of the logic involved in any automatic processing”

Trade secrets
- algorithms, method used
- weightings or criteria
- code
- software
- information about reference groups

Information about systems functionality
Inferences: assessments & predictions

- Assessments and predictions are often not verifiable and thus, difficult to correct or contest
A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI

*Columbia Business Law Review, Forthcoming*

85 Pages  •  Posted: 5 Oct 2018  •  Last revised: 14 Oct 2018

*Sandra Wachter*
University of Oxford - Oxford Internet Institute

*Brent Mittelstadt*
University of Oxford - Oxford Internet Institute

Date Written: September 13, 2018

**Abstract**

Big Data analytics and artificial intelligence (AI) draw non-intuitive and unverifiable inferences and predictions about the behaviors, preferences, and private lives of individuals. These inferences draw on highly diverse and feature-rich data of unpredictable value, and create new opportunities for discriminatory, biased, and invasive decision-making. Concerns about algorithmic accountability are often actually concerns about the way in which these technologies draw privacy invasive and non-verifiable inferences about us that we cannot predict, understand, or refute.
A right to reasonable inferences

**High risk inferences** - have low verifiability (e.g. predictive or opinion-based) and used for important decisions

Data, method for inferences (assessments & predictions)

- Statistical reliability
- Fairness
- Explanations
- Transparency (trade secrets, IP laws)
The future of AI requires dialogue between developers and society about not only what is possible, but also what is reasonable.
Thank you!

Dr. Sandra Wachter

sandra.wachter@oii.ox.ac.uk

@SandraWachter5