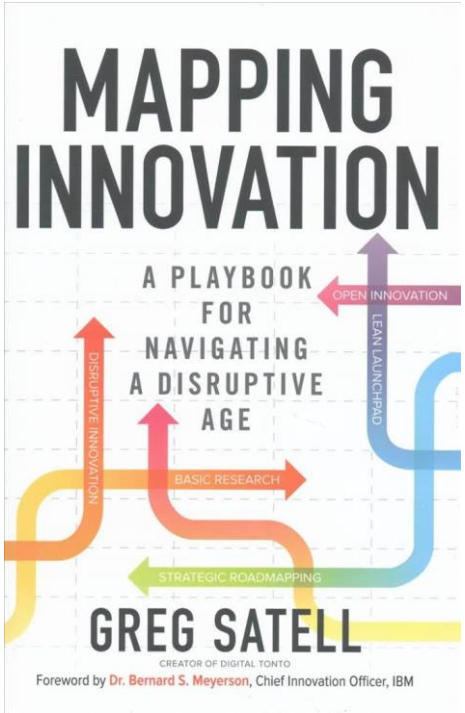
Fairness on Financial AI

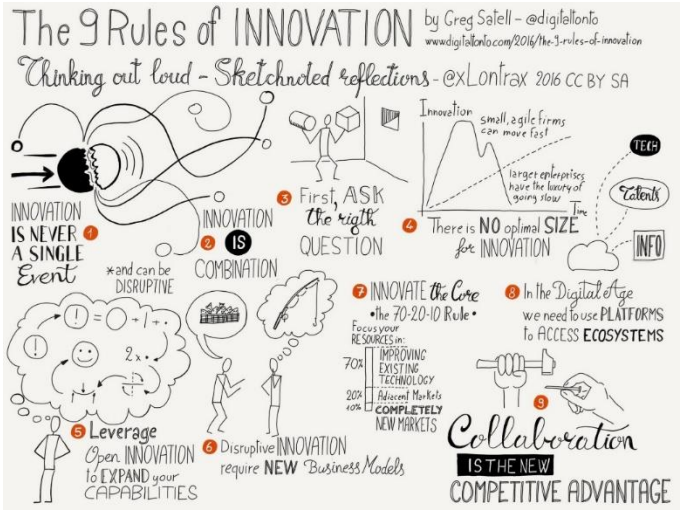
Experian Datalabs Uk & EMEA

Javier Campos
Head of Experian UK&I and EMEA Data Labs

Experian's approach to AI - Datalabs

Established in 2011 and with 4 global hubs – San Diego, London, Sao Paolo and Singapore, Experian Datalabs are constantly exploring the art of the possible, using a combination of scientists, technology and business focus.





Background

Fairness has been studied since the beginning of civilisation - whenever a decision to allocated resources needs to be done

Machine Learning, Data Science, Statistics

Law, policy, Advocacy (PSD2, GDPR, Equality act 2010 ..)

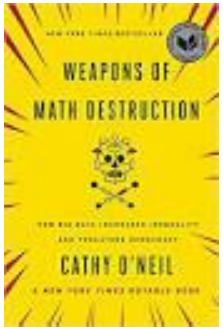
Empirical, social science, quantitative public policy

Applied math, CS theory, operations research, mathematical soc

Ethics, Philosophy, Qualitative Soc

Fairness – why should we care?

1. This is already happening in many places …



AI technologies automate Intelligent decisions by adapting and learning from local data at scale

2. AI, the new electricity, will be ubiquitous …

- From central operations with humans-

- To distributed intelligent decisions -

COMPAS – US tool to predict recidivism

Hiring – Big Tech company cancel App

EGov –App to report Potholes

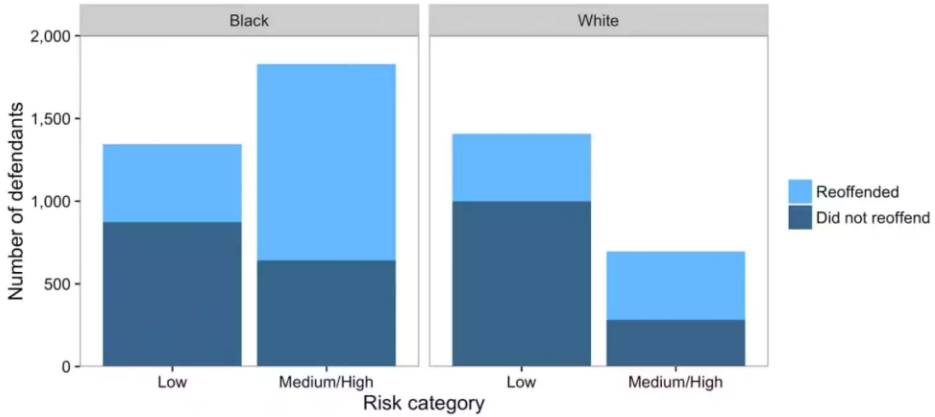
experian

Fairness is a complex issue: COMPAS

- **Northpointe's view** = the proportion of defendants who reoffend is approximately the same regardless of race.
- The overall recidivism rate for black defendants is higher than for white defendants (52 percent vs. 39 percent).
- Black defendants are more likely to be classified as medium or high risk (58 percent vs. 33 percent).
- **ProPublica's criticism** = Black defendants who don't reoffend are predicted to be riskier than white defendants who don't reoffend.

Key Point: Given the first 2 observations, the last two disparities are mathematically guaranteed!
So what can we do?

- It's hard to call a rule fair if it does not meet Northpointe's notion of fairness. A risk score of seven for black defendants should mean the same thing as a score of seven for white defendants.
- But we cannot disregard ProPublica's findings either, since classification errors disproportionately affect black defendants. Other policies might need to be explored.



Fairness is just a dimension of a full AI Framework

As the machines make more and more critical decisions, we need to consider a wider AI framework: Privacy, Explainability are closely related to Fairness and rest of areas to consider when creating AI models. The FAT/ML organisation was started in 2014 jointly by the machine learning and social science community.

F - FAIRNESS (models built are used to make unbiased decisions or predictions)

A - ACCOUNTABILITY (determine and assign responsibility for a judgement made by a machine)

C - CUSTOMER (Put the customer first)

T - TRANSPARENCY (being open and clear to the end user about how an outcome is made)

E - ETHICS (keeping in mind the ethical decisions that the machine might need to make)

S - SAFETY/SECURITY (ensuring the systems cause no harm and are secure against malicious behaviour)

Note these concepts are not orthogonal, for example both accountability and transparency can make systems fairer.

experian

Module: Fairness & Bias – Use Cases

Fairness & Bias

- I want to evaluate the bias/fairness of my already trained model.
- I have training data and I want to train a fair model.
- I have an existing model, which has been deemed to be unfair. How can I make it more fair?
- I am a modeller and need access to a cohesive and flexible toolbox for my modelling.

Normative Data	Vetted and Intuitive Fairness Metrics	Fair ML Algorithms	Fairness Modifiers
✓ ConsumerView ✓ Census ✓ CAIS/CATO/Bureaux ✓ Company House ✓ PSD2	✓ Group Fairness ✓ Individual Fairness	✓ Prioprietary ✓ Public	✓ Post-hoc Modelling ✓ Fairness Regularization ✓ Dataset Fairness Modifiers

Experian provides a unique combination of decisioning platform, normative data, and state-of-the-art algorithmic expertise.

experian

Module: Fairness & Bias – Feature Diagram

- What is my model doing?
- Is it accidently using variables it shouldn't?
- What is an appropriate measure of fairness?
- How do I evaluate models in line with my companies ethics and regulatory obligations?
- How do I make my current decisioning scores fairer?
- How do I build fairer models in the future?

Module Components

python™

Explainability	Measuring Fairness	Fairness for Existing Models	Training Fair Models
☐ Shapley Values & LIME	☐ Group Fairness: Disparate Impact	☐ Fair Metamodelling: Post-processing of Scores/Labels	☐ Dataset Debiasing
☐ Surrogate Models	☐ Individual Fairness: Equality of Opportunity		☐ Public algorithms
☐ Partial Dependence Plots			☐ Proprietary algorithms

experian™

Normative Data

Sources

Normative data sources will vary by region and the ability to join the aggregated data sources to the provided training data may be difficult.

When evaluating fairness in machine learning, the main protected attributes of interest at this time include:

Variable	Source	Key	Aggregation
GENDER	Planet Presumer/Consumerview	Name/PII	Personal
ETHNICITY	Census	Postcode	Output Areas*
RELIGION	Census	Postcode	Output Areas*
DISABILITY	Census	Postcode	Output Areas*
AGE	Consumerview/ Delphi/Client data	PII	Personal
COUNTRY OF BIRTH	Census	Postcode	Postcode
HOUSEHOLD COMPOSITION	Census/Consumerview	Postcode/PII	Output Areas*/Personal

Why?

Detect unfairness through missing variables at:

- **Person/Individual Level**
- **Aggregated level**

A Mathematical Representation (1 / 2)

By Friedler, Scheidegger, Venkatasubramanian

Algorithmic decision-making can be seen as a set of mappings between three spaces:

Decision space	Construct space	Observed space
Performance in college	Intelligence	IQ
Performance in college	Success in High School	GPA
Recividism	Propensity to commit crime	Family history of crime
Recidivism	Risk-averseness	Age
Employee Productivity	Knowledge of job	Number of Years of Experience

- The desired outcome is a mapping from CS to DS via some unknown complex function *o* of features that lie in the CS.
- In order to implement an algorithm that predicts the desired outcome, one must first extract usable data from CS, this is a collection of mappings from CS to OS.
- Machine Learning designs algorithms that approximate the mapping from CS to DS by a mapping ô from OS to DS.
- **The hope: ô = o**

A Mathematical Representation (2 / 2)

"Worldviews"

The existence of the observed space and the fact that the construct space is unobservable complicates data-driven decisions. In order to address this, we have to make assumptions about the construct space or the mapping between the construct space and the observed space.

Possible Assumptions:

- **What you see is what you get** – CS and OS are essentially the same
- **We are all the same** - If we assume there is some Structural Bias, i.e. existence of more distortion between groups than there is within groups when mapping between CS and OS, then we can assume that all groups look essentially the same. In other words, there are no inert differences between groups of individuals defined via certain discriminatory characteristics. (Alternatively interpretations, the groups are not equal but for the purposes of the decision-making process should be treated as if they were.)

The assumption can be a strategic decision, but this choice is critical to the decision-making process. The assumption determines what fairness means.

Allows for Individual fairness - asserts that decision making must treat people that are close in the observed space similarly and people that are far apart, differently.

Allows for Group fairness - asserts that the decision making should treat all groups in the same way

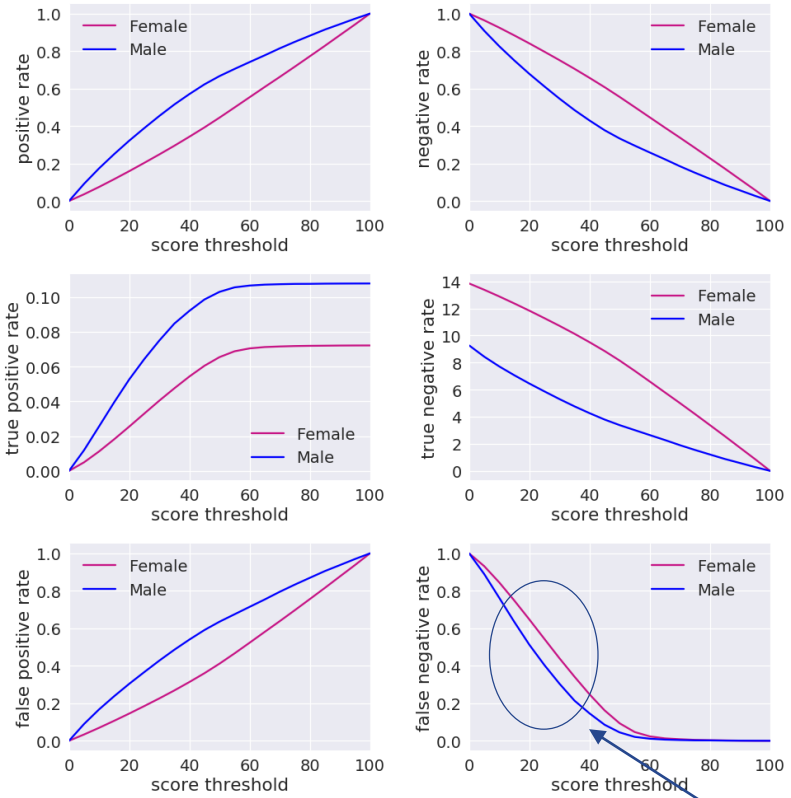
Ultimately, group and individual fairness are only compatible in specific circumstances.

experian

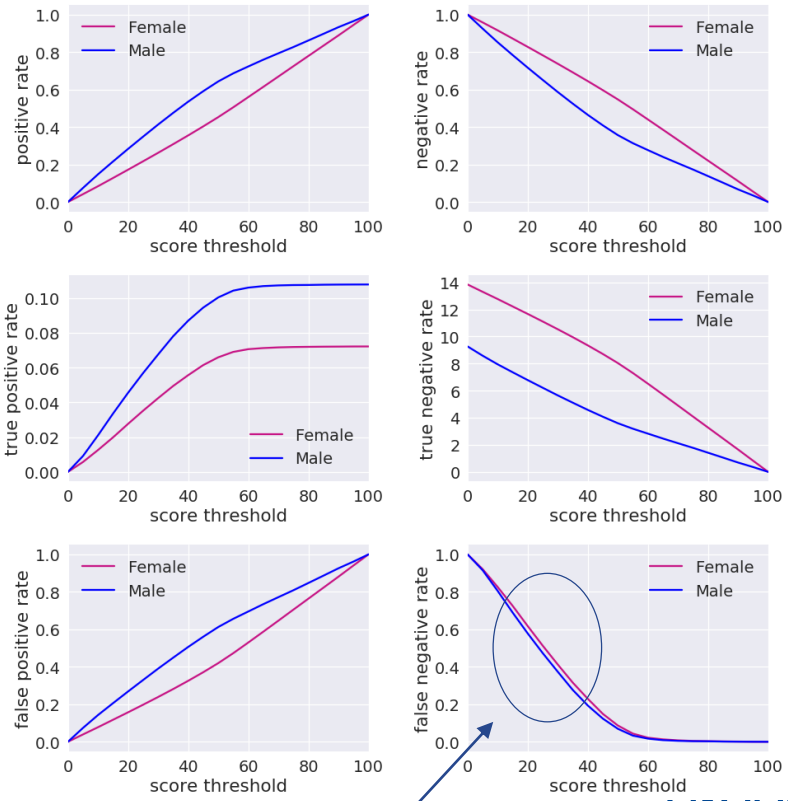
Example

Fraud Model and Gender

Model included title as a feature



Model without title as a feature

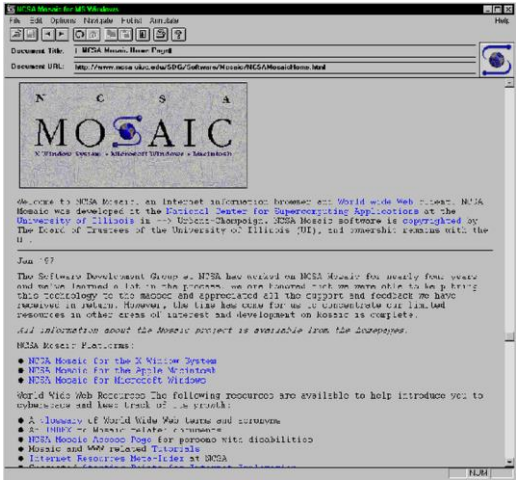


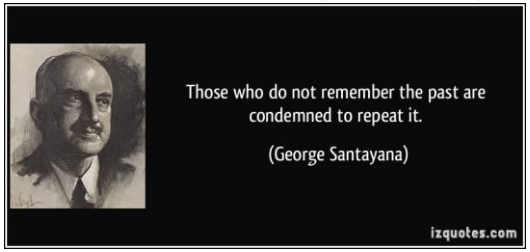
This model is biased against men

experian

The key Take aways…

Take away Point 1: When are we in the AI journey?





Software Engineering key stages

- 1945 to 1965: The origins
- **1965 to 1985: The software crisis**
- 1985 to 1989: "No Silver Bullet“ Software projects
- 1990 to 1999: Prominence of the Internet
- 2000 to Now: Lightweight methodologies

Software crisis(1965 – 1985):

- Many projects *ran over budget and schedule*. Some caused property damage. A few caused loss of life. Due to: productivity, quality and lack of qualified programmers

Are we about to enter a “**Data Crisis stage” in the AI journey?**

- Large gap on skillsets
- Impact on fairness – Fake news, violent use of Media
- Few deaths Tesla autopilot, industrial robots
- AI goes through hype periods followed by ‘Winters’ After the research peak in Deep Learning, a new generation of new algorithms will come which will make current platforms obsolete

There is no such a thing as a silver bullet AI platform. All giant Tech firms have resorted to custom builds but this requires very large pockets. Focus at this stage on specific short term business value, do not worry too much about standardization - just yet …

experian

Take away point 2: Fairness and Governance framework:

"A.I. makes mistakes as humans do, only faster and at scale"

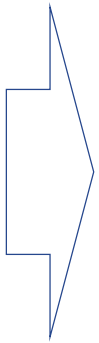
We can see a trend to start considering **Ethics as a quantifiable** (and measurable !) **business consideration** rather than as the current corporate social responsibility "soft" approach

As the machines make more and more critical decisions, we need to consider a wider AI framework: Privacy, Explainability are closely related to Fairness and rest of areas to consider when creating AI models. The regulators will go into this area so better prepare the organization.

experian

Take away Point 3: Lessons Learned on Fairness

- Fairness in machine learning is **a complex concept** that does not have a single solution. Simple ML model exhibit it, and Deep Learning models add the complexity of Explicability
- It is **very contextual**, the metrics to be used will depend on the context of the problem being solved.
- **Not a one-off task**, needs to be taken into account as an overall framework across everything to do with data – notably Privacy and Explainability - and across the end to end lifecycle of the algorithm.

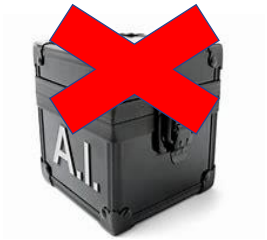


- Ensure your team understand the complexity, seek collaboration from other fields, Open sources and specialised parties with the right skillset
- In many cases, it is easier to detect than to fix it.
- As well as the algorithm, the data is extremely important (e.g Boeing 737 Max 8)

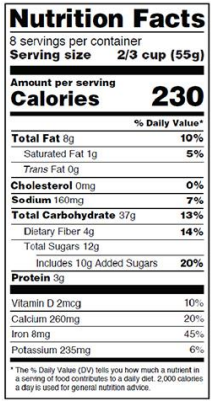
experian

Future of AI… Trusted AI

Several dimensions need to be address when designing an AI system – a framework for trusted AI:



- **F - FAIRNESS & ETHICS**
- **A - ACCOUNTABILITY**
- **C - CUSTOMER**
- **T - TRANSPARENCY**
- S - SAFETY/SECURITY



Nutrition Facts

8 servings per container
Serving size **2/3 cup (55g)**

Amount per serving
Calories **230**

	% Daily Value*
Total Fat 8g	**10%**
Saturated Fat 1g	**5%**
Trans Fat 0g	
Cholesterol 0mg	**0%**
Sodium 160mg	**7%**
Total Carbohydrate 37g	**13%**
Dietary Fiber 4g	**14%**
Total Sugars 12g	
Includes 10g Added Sugars	**20%**
Protein 3g	
Vitamin D 2mcg	10%
Calcium 260mg	20%
Iron 8mg	45%
Potassium 235mg	6%

* The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

Data Science Facts

Fairness

1. Was the dataset and model checked for biases?
2. Was any bias mitigation performed on the dataset?

Accountability

1. Does the algorithm has a clear definition of who is accountable for all their outcomes – direct and indirect?

Customer

1. Was the service checked for robustness against adversarial attacks?
2. Is usage data from service operations retained/stored/kept?
3. What will be expected behavior if the input deviates from training/testing data?
4. What kind of governance is employed to track the overall workflow of data to AI service?

Transparency

1. Are algorithm outputs explainable/interpretable
2. Who is the target user of the explanation (ML expert, domain expert, general consumer, regulator, etc.)
3. Was the service tested on any additional datasets? Do they have a datasheet or data statement?


experian™