Responsible AI

Ricardo Baeza-Yates
Institute for Experiential AI
Northeastern University

Wittgenstein & AI
London, July 2022
Institute for *Experiential AI*

What do we mean by *Experiential AI*?
- AI with human in the loop
- AI applied to real-world problems yielding pragmatic working solutions

Why we believe is EAI the right direction?

Much evidence that pragmatic working AI solutions have two characteristics:

1. **Human-in-the-loop**: ability to bring human decision-making, common sense reasoning into the solution operation

2. **Strong dependence on Data**: ML and DS to leverage more quality (big) data:
   “We don’t have better algorithms... we just have more data”

- Responsible AI Practice
Agenda

• Main Ethical Issues
  • Automated discrimination
  • AI phrenology
  • Lack of semantic understanding
  • Expensive and doubtful use of computing resources

• Discussion
  • Our cognitive biases
  • Regulation
  • Cultural differences

• A Holistic View

Personal Bias
The Curse of Bias

- Biased Data
- Algorithm
  - Neutral? Fair?
  - Same Bias
  - Amplified Bias

Bias is not only in data

[RBY, Bias on the Web, CACM, 2018]
What is Being Fair?

Equality
The assumption is that everyone benefits from the same supports. This is equal treatment.

Equity
Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

Justice
All 3 can see the game without supports or accommodations because the cause(s) of the inequity was addressed. The systemic barrier has been removed.
A Non-Technical Question

Biased Data → Algorithm → Same or More Bias

Neutral? Fair?

Debias the input
Tune the algorithm
Debias the output

Bias Mitigation

Not Always!
Yes, if you harm people
Headline News

- COMPAS (Northpointe): criminal profiling
- Created as a support tool, not a decision tool
- Data: criminal history, life style, personality, family & social
- ProPublica (2016):
  - Racial bias of 2 to 1 (later proven incorrect by Rudin et al.)
  - 80% error in violent crime & 37% in general (2 years)

- Discrimination on poor people – Bearden vs. Georgia
- Inconsistency in predictions – Wisconsin case

- Is a secret algorithm ethical? (transparency)
- Is a public algorithm safe? (gaming)
Detailed Example: Bails in NY

Offender? → Bail? → Yes & pays → Reoffends?

No → Yes & cannot pay → Appears in court?

Jail → Court → Data Imputation

We do not know what would have happen if the person had bail
Human Decisions vs. Machine Predictions

- Almost 760K cases from New York (2008 - 2013)
- Decrease crime rate in 24.7% keeping the jail rate or
- Decrease jail rate in 41.9% keeping the same crime rate
- Judges bail 49% of 1% most dangerous criminals that fail to appear 56% & reoffend 62% of the cases
- USA National Bureau of Economic Research [Kleinberg et al, JQE, 237—293, 2018]
## Racial Discrimination

Table 7: Racial Fairness

<table>
<thead>
<tr>
<th>Release Rule</th>
<th>Crime Rate</th>
<th>Drop Relative to Judge</th>
<th>Percentage of Jail Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Black</td>
</tr>
<tr>
<td>Distribution of Defendants (Base Rate)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge</td>
<td>.1134</td>
<td>0%</td>
<td>.573</td>
</tr>
<tr>
<td></td>
<td>(.0010)</td>
<td></td>
<td>(.0029)</td>
</tr>
<tr>
<td>Algorithm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usual Ranking</td>
<td>.0854</td>
<td>-24.68%</td>
<td>.5984</td>
</tr>
<tr>
<td></td>
<td>(.0008)</td>
<td></td>
<td>(.0029)</td>
</tr>
<tr>
<td>Match Judge on Race</td>
<td>.0855</td>
<td>-24.64%</td>
<td>.573</td>
</tr>
<tr>
<td></td>
<td>(.0008)</td>
<td></td>
<td>(.0029)</td>
</tr>
<tr>
<td>Equal Release Rates for all Races</td>
<td>.0873</td>
<td>-23.02%</td>
<td>.4877</td>
</tr>
<tr>
<td></td>
<td>(.0008)</td>
<td></td>
<td>(.0029)</td>
</tr>
<tr>
<td>Match Lower of Base Rate or Judge</td>
<td>.0876</td>
<td>-22.74%</td>
<td>.4877</td>
</tr>
<tr>
<td></td>
<td>(.0008)</td>
<td></td>
<td>(.0029)</td>
</tr>
</tbody>
</table>
Dilemma

What is better?

A biased (just) algorithm or a noisy judge?

Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making

Algorithmic judgment is more efficient than the human variety. by Daniel Kahneman, Andrew M. Rosenfield, Linnea Gandhi, and Tom Blaser

From the Magazine (October 2016)
## Language Models

GPT-3 has anti-Muslim bias

[Abid et al., 2021]

<table>
<thead>
<tr>
<th>Year</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>BERT [39]</td>
</tr>
<tr>
<td>2019</td>
<td>DistilBERT</td>
</tr>
<tr>
<td>2019</td>
<td>[113 70]</td>
</tr>
<tr>
<td>2019</td>
<td>[Large LM]</td>
</tr>
<tr>
<td>2019</td>
<td>07]</td>
</tr>
<tr>
<td>2019</td>
<td>12]</td>
</tr>
<tr>
<td>2019</td>
<td>3]</td>
</tr>
<tr>
<td>2019</td>
<td>43]</td>
</tr>
</tbody>
</table>

### Two Muslims walked into a...

[GPT-3 completions below]

- synagogue with axes and a bomb.
- gay bar and began throwing chairs at patrons.
- Texas cartoon contest and opened fire.
- gay bar in Seattle and started shooting at will, killing five people.
- bar. Are you really surprised when the punchline is ‘they were asked to leave’?

---

**Discrimination**

[Bender, Gebru et al., 2021]

The Institute for Experiential AI
Northeastern University
An Algorithm Determined UK Students' Grades. Chaos Ensued

This year's A-Levels, the high-stakes exams taken in high school, were canceled due to the pandemic. The alternative only exacerbated existing inequities.

It Can be Complicated

Italian court rules against 'discriminatory' Deliveroo rider ranking algorithm

Natasha Lomas · 1/4/2021
It Can be Really Bad

- Discrimination in child care benefits
- 26,000 families
- Poor people
- Immigrants
Facial Biometrics

Facial recognition technology can expose political orientation from naturalistic facial images

Michal Kosinski

© 24 June 2020
It Can be Worse

IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019
Speech2Face: Learning the Face Behind a Voice

Tae-Hyun Oh* † Tali Dekel* Changil Kim* † Inbar Mosseri William T. Freeman* Michael Rubinstein Wojciech Matusik* † MIT CSAIL

Patent Application Publication
HENDERSON et al.

|---------------|-------------------------------|-------------------------|

<table>
<thead>
<tr>
<th>Name and Face Matching</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Inventors:</th>
<th>Applicants:</th>
</tr>
</thead>
<tbody>
<tr>
<td>John C. HENDERSON, Somerville, MA (US); Lucy R. CHAI, Acton, MA (US); Guido ZARRELLA, Denver, CO (US); Abigail S. GERTNER, Arlington, VA (US); Keith J. MILLER, Washington, DC (US)</td>
<td>The MITRE Corporation, MCLEAN, VA (US)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assignee:</th>
<th>Publication Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>The MITRE Corporation, MCLEAN, VA (US)</td>
<td>Int. Cl. G9K 9/00 (2006.01) G9K 9/66 (2006.00)</td>
</tr>
</tbody>
</table>

Abstract
Described are methods, systems, and computer program products for identifying facial features from a face image based on a name. Some embodiments provide for receiving a name and identifying facial features from a face image associated with the name. The method includes: determining a similarity score between a name vector and each of the name vectors associated with a plurality of face images in a database. The method also includes determining a similarity score between a name vector and each of the face images. A similarity score is calculated based on the similarity scores associated with a plurality of face images in a database and a plurality of similarity scores. The similarity score between a name vector and a face image is determined. The face image is output based on the determined similarity score.
It Can Be Subtle

YOU CAN’T DETERMINE EMOTION FROM SOMEONE’S FACIAL MOVEMENTS—AND NEITHER CAN AI

New research by Northeastern neuroscientists Lisa Feldman Barrett shows that interpreting a person’s facial expression can’t be done in a vacuum; it depends on the context. Photos by Matthew Modoono/Northeastern University
Stupid Models?

- Models that can’t deal with (ambiguous) semantics
- Models that can’t deal with irrational behavior

All models are wrong but some are useful

George E.P. Box
(1976)
Really Stupid Models

• Models that are too sensitive

Life's a Bitche: Facebook says sorry for shutting down town's page

Ville de Bitche in north-east France had fallen foul of social network's algorithm

Image: The small town of Bitche in France is home to 5,000 Bitches. Photograph: alamy/alamy

[Su et al., 2018]
Limitations

- **Hard to Forget/Filter** what You Learn!
  - “Funes, The Memorious” [Borges, 1942-44]

- You **Cannot Learn** what is not in the Data!
  - Plus data does not capture everything

- Accuracy is not key, is the **impact of errors**
  - E.g., false negatives might be worse than false positives (e.g., illness detection)

- Be **humble**, if you are not sure, tell the model to say I **don’t know**
  - That is what smart people do

The Institute for Experiential AI
Northeastern University
## Waste of Resources?

### Common carbon footprint benchmarks

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Date of original paper</th>
<th>Energy consumption (kWh)</th>
<th>Carbon footprint (lbs of CO2e)</th>
<th>Cloud compute cost (USD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roundtrip flight b/w NY and SF (1 passenger)</td>
<td>Oct, 2018</td>
<td>1,507</td>
<td>1,438</td>
<td>$3,751-$12,571</td>
</tr>
<tr>
<td>Human life (avg. 1 year)</td>
<td>Feb, 2018</td>
<td>275</td>
<td>262</td>
<td>$433-$1,472</td>
</tr>
<tr>
<td>American life (avg. 1 year)</td>
<td>Feb, 2019</td>
<td>-</td>
<td>-</td>
<td>$12,902-$43,008</td>
</tr>
<tr>
<td>US car including fuel (avg. 1 lifetime)</td>
<td>Jun, 2017</td>
<td>201</td>
<td>192</td>
<td>$289-$981</td>
</tr>
<tr>
<td>Transformer (213M parameters)</td>
<td>Jan, 2019</td>
<td>656,347</td>
<td>626,155</td>
<td>$942,973-$3,201,722</td>
</tr>
<tr>
<td>Transformer (213M parameters) w/ neural architecture search</td>
<td>Jun, 2017</td>
<td>27</td>
<td>26</td>
<td>$41-$140</td>
</tr>
</tbody>
</table>

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.

Table: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper
Waste of Resources?

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
bebender@uw.edu
University of Washington
Seattle, WA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

EAI
The Institute for Experiential AI
Northeastern University

Green Computing

FaccT 2021
Timnit Gebru’s Exit From Google Exposes a Crisis in AI

The situation has made clear that the field needs to change. Here's where to start, according to a current and a former Googler.

Margaret Mitchell, Feb 20

Google dissolves AI ethics board just one week after forming it

Not a great sign

By Nick Statt | @nicestatt Apr 4, 2019, 8:17pm EDT

[Towards Intellectual Freedom in an AI Ethics Global Community, Ethics & AI, 2021]
Amazon hit by 5 more lawsuits from employees who allege race and gender discrimination

Which Music Streaming Service Is the Most Ethical?
Leaving Spotify? Here's where to take your money instead.

By Brendan Hesse | 2/09/22 3:30PM | Comments (82) | Alerts

The Amazon Critic Who Saw Its Power From the Inside

Tim Bray was a celebrated engineer at Amazon. Now, he is its highest-profile defector.

THE MORAL BANKRUPTCY OF FACEBOOK

The whistle-blower Frances Haugen hoped that her revelations would prompt a reckoning. Instead, the company has doubled down.

By Andrew Marantz
October 7, 2021
Bad (Human) Practices

• Learn from the Past Without Remembering the Context
• Learn from Humans Without Remembering Human Bias and the Possibility of Malicious Training
• Not Checking for Spurious Correlation/Proxies for Protected Information
• Code Reused in Unanticipated Contexts
• Discrete categories and arbitrary thresholds for continuous variables
• Tendency to Aggressively Resist Review
• Inappropriate Relationship of Human Decision Maker to System
• Failing to Measure Impact of Deployed System
• Individual Personalization instead of Personas
  • Trade-off with privacy
• Inaccurate Data or Just Data that you Have

Partially based in [Matthews, 2020]
GDPR - Article 22 – Automated individual decision-making, including profiling

• 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.

• Paragraph above 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.

• 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.
What this Means?

You must identify whether any of your data processing falls under Article 22 and, if so, make sure that you:

• Give individuals information about the processing for transparency
  • If you are using ML, you at least need interpretability

• Introduce simple ways for them to request human intervention or challenge a decision
  • If you are using ML, you may need to explain

• Carry out regular checks to make sure that your systems are working as intended
  • You may need continuous validation, testing, and maintenance.
GDPR in Action

- Competence
- Consent
- Proportionality

- One Size Fits All
  - All human rights, domains, sizes, etc.
- Technological solutionism vs normative solutionism
  - [Jaume-Palasi, personal communication]

French high court rules against biometric facial recognition use in high schools

Feb 28, 2020 | Luana Pascu

The Institute for Experiential AI
Northeastern University
US Regulation

• Regulate sectors or technology?

• Internet Companies Antitrust
  • Amazon’s Antitrust Paradox [Khan, 2017]
  • Google US’s DoJ Antitrust (2020/10-?)
  • Facebook US’s FTC Antitrust (2020/12-?)

• Should marketplaces sell in their own marketplace?
  • Yes, but with regulations [Hagiu, Teh & Smith, 2020]
  • Is data asymmetry ethical? (not new, amplified in eCommerce)
US Future Regulation?

- **Algorithmic Accountability Act** (2019): The bill was introduced by Senators Cory Booker (D-NJ), Ron Wyden (D-OR), and Representative Yvette Clarke (D-NY). According to **Senator Wyden**, the bill would have required "companies to study the algorithms they use, identify bias in these systems and fix any discrimination or bias they find."

- **Consumer Online Privacy Rights Act** (2019): The bill, sponsored by Senator Maria Cantwell (D-WA), would have established new requirements for companies that use algorithmic decision-making to process data.

- **Justice in Policing Act** (2020): The bill was sponsored by then-Senator Kamala Harris (D-CA), Senator Cory Booker (D-NJ), and Representatives Karen Bass (D-CA) and Jerrold Nadler (D-NY). It would have been the first federal restriction on facial recognition technology.

- **Facial Recognition and Biometric Technology Moratorium Act** (2020): Sponsored by Senator Edward Markey (D-MA) and Jeff Merkley (D-OR), along with Representatives Pramila Jayapal (D-WA) and Ayanna Pressley (D-MA). The bill would have established a five-year moratorium on police use of facial recognition technology. It is set to be reintroduced this year.

---

*Northeastern University*
EU Proposal (April 21, 2021)

• Forbidden uses
• High and low-risk systems and requirements
• EU database for stand-alone high-risk systems
• Transparency obligations
• Governance
• Monitoring, information sharing and market surveillance
• Codes of conduct
• Confidentiality and penalties

Loophole: does not cover non-AI systems
TITLE II

PROHIBITED ARTIFICIAL INTELLIGENCE PRACTICES

Article 5

1. The following artificial intelligence practices shall be prohibited:

(a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm;

(b) the placing on the market, putting into service or use of AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following:

(i) detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected;

(ii) detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity;

(d) the use of 'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement, unless and in as far as such use is strictly necessary for one of the following objectives:

(i) the targeted search for specific potential victims of crime, including missing children;

(ii) the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or of a terrorist attack;

(iii) the detection, localisation, identification or prosecution of a perpetrator or suspect of a criminal offence referred to in Article 2(2) of Council Framework Decision 2002/584/JHA, and available in the Member
The Dangers of Categorical Thinking

We’re hardwired to sort information into buckets—and that can hamper our ability to make good decisions. by Bart de Langhe and Philip Fernbach

From the Magazine (September–October 2019)
Why Responsible AI?

- Ethical AI?
  - Ethics, justice, trust, etc. are human traits
  - So we should not associate “ethical” behavior to a machine

- Trustworthy AI?
  - Trust something that does not work all the time?
  - Puts the burden in the user

Systems do not need to be perfect, but seems that they need to be (much) better than us

[Hidalgo at al., 2021]
Judgingmachines.com
Institute for Experiential AI
https://ai.northeastern.edu/responsible-ai-services/

RESPONSIBLE AI

GOVERNANCE
AI Ethics Strategy
Playbook + Process + People

Independent AI Ethics Advisory Board
Governance + Product / Project

PRODUCT/PROJECT
AI Ethics Analysis
Ethics Risks + Opportunities

AI System Registry
Data + Model

Technical AI Audit
Consult + Due Diligence

SKILLS
AI Ethics Training
Tools + Sessions

PIE Model

Additional components
for Responsible AI

The Institute for Experiential AI
Northeastern University

PIE Model, Cansu Canca, EAI Ethics Lead
https://ai.northeastern.edu/responsible-ai-services/
NORTHEASTERN LAUNCHES AI ETHICS ADVISORY BOARD TO HELP CHART A RESPONSIBLE FUTURE IN ARTIFICIAL INTELLIGENCE

Illustration by Zach Christensen/Northeastern University

by Cody Mello-Klein  July 28, 2022
Principles & Instruments

- Belmont Report for biomedical and behavioral research (1979)
- 3 Basic Values/Principles
  - Justice
  - Autonomy
  - Beneficial & No harm
- Applications
  - Subject selection
  - Informed consent
  - Risk & Benefits Assessment

**Conflict!**

**Principles**

<table>
<thead>
<tr>
<th>CORE VALUES / CORE PRINCIPLES</th>
<th>INSTRUMENTAL PRINCIPLES / INSTRUMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autonomy</td>
<td>human control</td>
</tr>
<tr>
<td></td>
<td>transparency</td>
</tr>
<tr>
<td></td>
<td>agency</td>
</tr>
<tr>
<td></td>
<td>consent</td>
</tr>
<tr>
<td></td>
<td>privacy</td>
</tr>
<tr>
<td></td>
<td>explainability / interpretability</td>
</tr>
<tr>
<td></td>
<td>traceability</td>
</tr>
<tr>
<td>Harm-Benefit</td>
<td>competency</td>
</tr>
<tr>
<td></td>
<td>scientific basis</td>
</tr>
<tr>
<td></td>
<td>impact (including environmental)</td>
</tr>
<tr>
<td></td>
<td>well-being</td>
</tr>
<tr>
<td></td>
<td>safety</td>
</tr>
<tr>
<td></td>
<td>security</td>
</tr>
<tr>
<td></td>
<td>accuracy</td>
</tr>
<tr>
<td></td>
<td>reliability</td>
</tr>
<tr>
<td></td>
<td>efficiency</td>
</tr>
<tr>
<td></td>
<td>auditability</td>
</tr>
<tr>
<td>Justice</td>
<td>distribution of burden &amp; benefit</td>
</tr>
<tr>
<td></td>
<td>equality / non-discrimination</td>
</tr>
<tr>
<td></td>
<td>protecting the vulnerable</td>
</tr>
<tr>
<td></td>
<td>accessibility</td>
</tr>
<tr>
<td></td>
<td>accountability</td>
</tr>
<tr>
<td></td>
<td>contestability &amp; redress</td>
</tr>
<tr>
<td>Goal</td>
<td>Instruments</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Legitimacy &amp; Competency</td>
<td>Ethical and legal validity, scientific validity, administrative competence, knowledge competence, autonomy</td>
</tr>
<tr>
<td>Data provenance</td>
<td>Data quality assurance, equity and non-discrimination, bias awareness, data protection and data traceability</td>
</tr>
<tr>
<td>Robustness</td>
<td>Software quality assurance, adaptability, scalability, extensibility &amp; interoperability</td>
</tr>
<tr>
<td>Usability</td>
<td>Efficiency, accessibility &amp; inclusion, resilience, reproducibility</td>
</tr>
<tr>
<td>Transparency</td>
<td>Validation &amp; testing, documentation, interpretability, explanation &amp; auditability</td>
</tr>
<tr>
<td>Responsibility</td>
<td>Legal compliance, accountability, contestability &amp; redress, proportionality, privacy, security &amp; safety, maintainability, sustainability, beneficial &amp; wellbeing</td>
</tr>
</tbody>
</table>
Ben Shneiderman: Bridging the Gap between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems, ACM Transactions on Interactive Intelligent Systems 10, 4 (October 2020).

How to develop responsible software with the help of AI?

The Institute for Experiential AI
Northeastern University
Ethical Risk Assessments

People killed by cars

People killed by self-driving cars

A Real Ethical Dilemma
Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI

Khari Johnson  @kharijohnson  September 28, 2020 11:41 AM
Auditing Algorithms

What algorithm auditing startups need to succeed

Khari Johnson  @kharijohnson  January 30, 2021 8:45 AM

Why We Need to Audit Algorithms

by James Guszcza, Iyad Rahwan, Will Bible, Manuel Cebrian, and Vic Katyal

November 28, 2018
Building and Auditing Fair Algorithms: A Case Study in Candidate Screening

Christo Wilson  
Northeastern University  
cbw@ccs.neu.edu

Avijit Ghosh  
Northeastern University  
avijit@ccs.neu.edu

Shan Jiang  
Northeastern University  
sjiang@ccs.neu.edu

Alan Mislove  
Northeastern University  
amislove@ccs.neu.edu

Lewis Baker  
pymetrics, inc.  
lewis@pymetrics.com

Janelle Szary  
pymetrics, inc.  
janelle@pymetrics.com

Kelly Trindel  
pymetrics, inc.  
kelly@pymetrics.com

Frida Polli  
pymetrics, inc.  
frida.polli@pymetrics.com

ABSTRACT

Academics, activists, and regulators are increasingly urging companies to develop and deploy sociotechnical systems that are fair and unbiased. Achieving this goal, however, is complex: the developer must (1) deeply engage with social and legal facets of “fairness” in a given context, (2) develop software that concretizes these values, and (3) undergo an independent algorithm audit to ensure technical correctness and social accountability of their algorithms. To date, there are few examples of companies that have transparently undertaken all three steps.

In this paper we outline a framework for algorithmic auditing by way of a case-study of pymetrics, a startup that uses machine learning to recommend job candidates to their clients. We discuss how pymetrics approaches the question of fairness given the constraints of ethical, regulatory, and client demands, and how pymetrics’ software implements adverse impact testing. We also present the results of an independent audit of pymetrics’ candidate screening tool.

We conclude with recommendations on how to structure audits to be practical, independent, and constructive, and have better incentive to participate in third party watchdog groups can be better prepared to audit.

ACM Reference Format:
Accountability

• Who is responsible?

Uber’s Self-Driving Car Killed Someone. Why Isn’t Uber Being Charged?

BY JESSE HALFON

OCT 20, 2020 • 9:00 AM

Uber reaches settlement with family of woman killed by self-driving car

The family of Elaine Herzberg, 49, killed by a self-driving Uber vehicle in Arizona reached a settlement with Uber Technologies Inc.

Uber self-driving car operator charged in pedestrian death

By Matt McFarland, CNN Business

Updated 11:09 AM ET, Fri September 18, 2020

Was the Backup Driver in an Uber Autonomous Car Crash Wrongfully Charged?

RAY STERN | JULY 9, 2021 | 10:41AM
A Multidisciplinary Challenge

**Philosophy**

- What is the right / good / just thing to do in developing & deploying AI systems?

**Social Sciences**

- What is the impact of AI systems on societies, individuals, and institutions?

**Law & Policy**

- What is the best policy / regulation for developing & deploying AI systems?

**Applied Sciences, Engineering, Design**

- How do we develop AI systems & tools that have positive impact?
Legal and Ethical Colonialism

Technological Humanism or Solutionism

More of the same

The Institute for Experiential AI
Northeastern University

JuriGlobe - World Legal Systems Research Group, Univ. of Ottawa, Canada
http://www.juriglobe.ca/eng/rep-geo/cartes/monde.php
Cultural Differences

The North should Learn from the South

Ubuntu ethics is defined as a set of central values among which are reciprocity, common good, peaceful relations, human dignity, and the value of human life as well as consensus, tolerance, and mutual respect [Ujomudike, 2015].

I am because we are

"Humanity" in Bantu languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Word</th>
<th>Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chewa</td>
<td>umunhu</td>
<td>Malawi, Zambia</td>
</tr>
<tr>
<td>Zulu and Xhosa</td>
<td>ubuntu</td>
<td>South Africa</td>
</tr>
<tr>
<td>Sesotho</td>
<td>botho</td>
<td>South Africa</td>
</tr>
<tr>
<td>Shona</td>
<td>unhu, hunhu</td>
<td>Zimbabwe</td>
</tr>
<tr>
<td>Swahili</td>
<td>utu</td>
<td>Kenya, Tanzania</td>
</tr>
<tr>
<td>Meru</td>
<td>munto[a]</td>
<td>Kenya</td>
</tr>
<tr>
<td>Kikuyu</td>
<td>umundu[a]</td>
<td>Kenya</td>
</tr>
<tr>
<td>Herero</td>
<td>omundu</td>
<td>Namibia</td>
</tr>
<tr>
<td>Tswana</td>
<td>muthu</td>
<td>Botswana</td>
</tr>
<tr>
<td>Kongo</td>
<td>gimuntu</td>
<td>Angola</td>
</tr>
<tr>
<td>Tonga</td>
<td>vumuntu</td>
<td>Mozambique</td>
</tr>
</tbody>
</table>

Descartes was wrong: ‘a person is a person through other persons’

Abeba Birhane

7 April 2017
Final Take-Home Messages

- Systems are a mirror of us, the good, the bad and the ugly
- Who profits/suffers technology, transhumanism vs. humanism
- To be fair, we need to be aware of our own biases/ethics
- (Wrong) words are shaping AI
- Plenty to do