by Dr. André dos **Santos**

[X]ai

ABOUT THIS COURSE

PART 1

- Bias
- Fairness

PART 2

- RAI Definitions
- XAI

PART 3

- Regulations
- Standards
- Guidelines

SETTING EXPECTATIONS

FIRST STEPS

RAI deployment is a progressive and evolving practice

FREE RESOURCES

Open source and freely available tools

ONGOING JOURNEY

Continuous effort is required to remain ethical and effective

NO ONE-SIZE-FITS-ALL

RAI solutions are highly domain dependent

PART 1

Bias & Fairness

LINKS

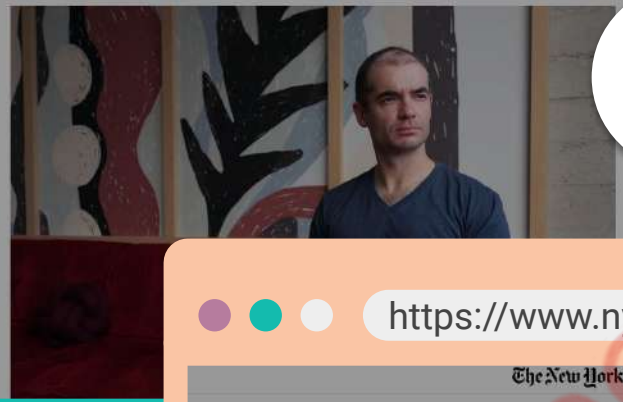
[X]ai

https://www.wired.com

WIRED SECURITY POLITICS GEAR BACKCHANNEL BUSINESS SCIENCE CULTURE IDEAS MERCH

OpenAI's Long-Term AI Risk Team Has Disbanded

The entire OpenAI team focused on the existential dangers of AI has either resigned or been absorbed into other research group



https://www.cnn.com

CNN Business Markets Tech Media Calculators Videos

S&P 500	5,303.27	0.12% ▲

Fear & Greed Index →

Greed
is driving the US market

Latest Market News →

Disneyland character performers vote t

Minnesota lawmakers strike minimum p

AI could pose 'extinction-level' threat to humans and the US must intervene, State Dept.-commissioned report warns

https://www.nytimes.com

The New York Times

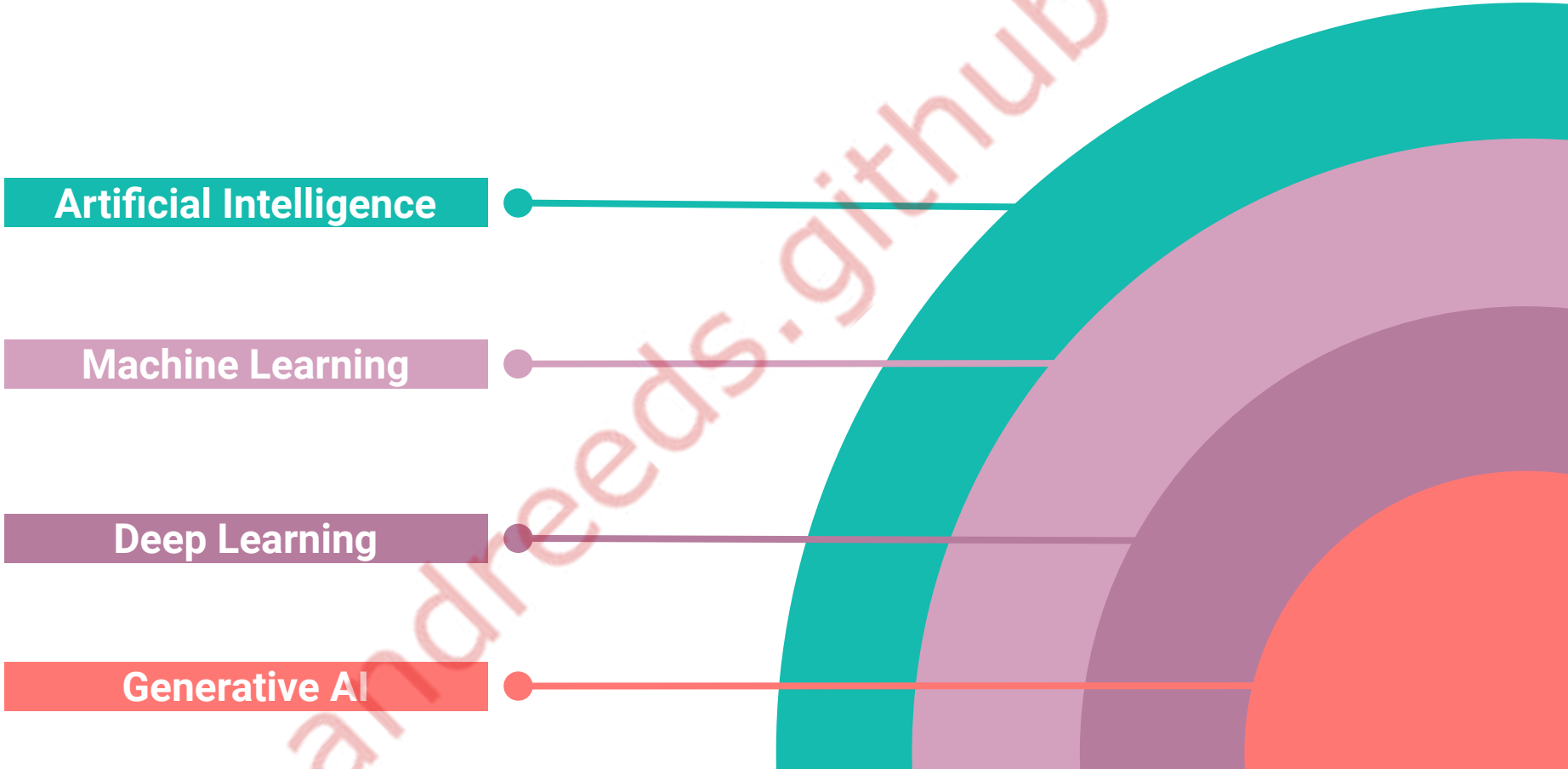
Artificial Intelligence › When A.I. Takes Your Voice Google's A.I. Evolution A.I.'s 'Her' Era Arrives OpenAI's Old-Fashioned Library Faces Q

How to Tell if Your A.I. Is Conscious

might indicate the presence of some presence in a machine.

Share full article 653

AI IS MORE THAN GENERATIVE AI



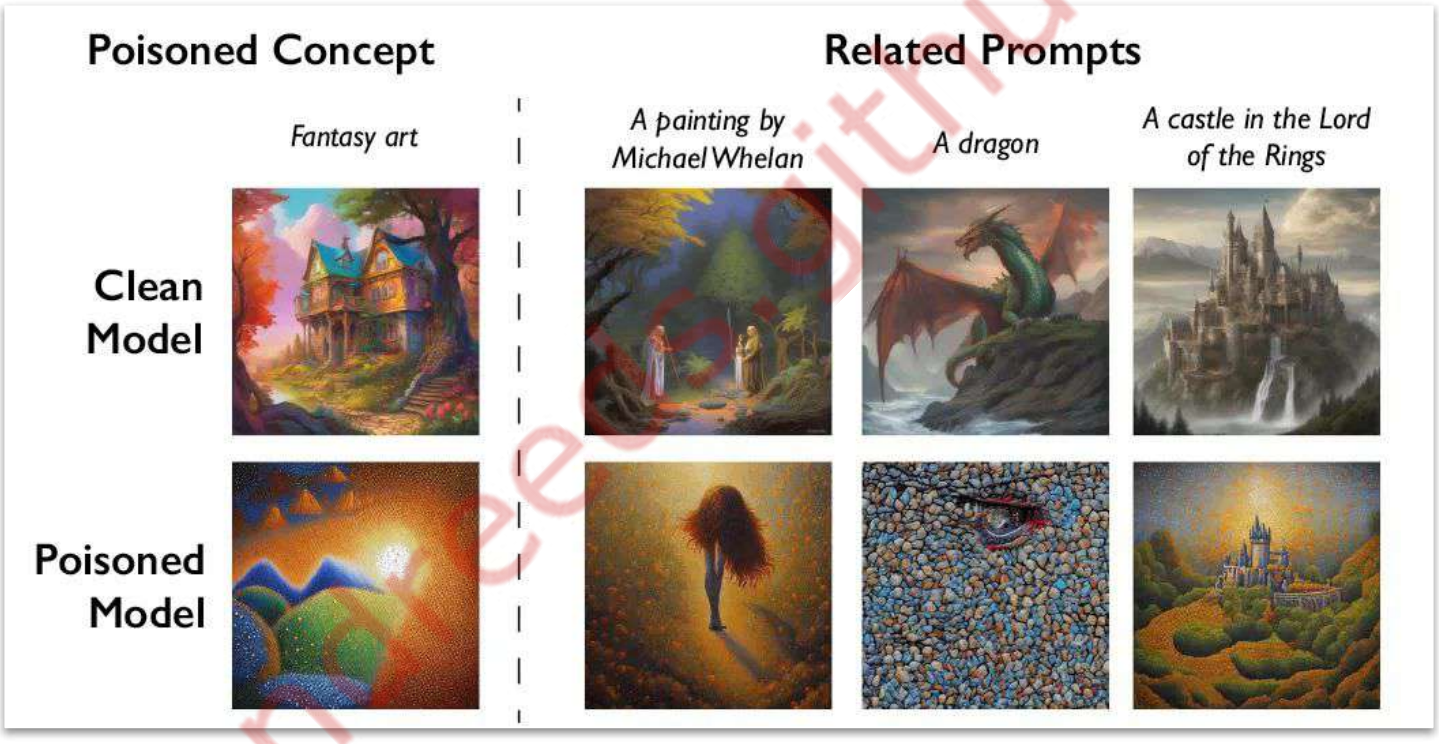
ARTIFICIAL INTELLIGENCE

This new data poisoning tool lets artists fight back against generative AI

The tool, called Nightshade, messes up training data in ways that could cause serious damage to image-generating AI models.

By Melissa Heikkilä October 23, 2023

NIGHTSHADE



IDENTIFYING BIAS

“AI bias is the phenomenon that occurs when an AI algorithm produces results that are systemically prejudiced due to erroneous assumptions in the machine learning process.”

—The Internet

AI bias is any systematic error that results in an unfair model

BIAS IN AI

DATA AGGREGATION

MODEL BUILDING & IMPLEMENTATION

BIAS IN AI

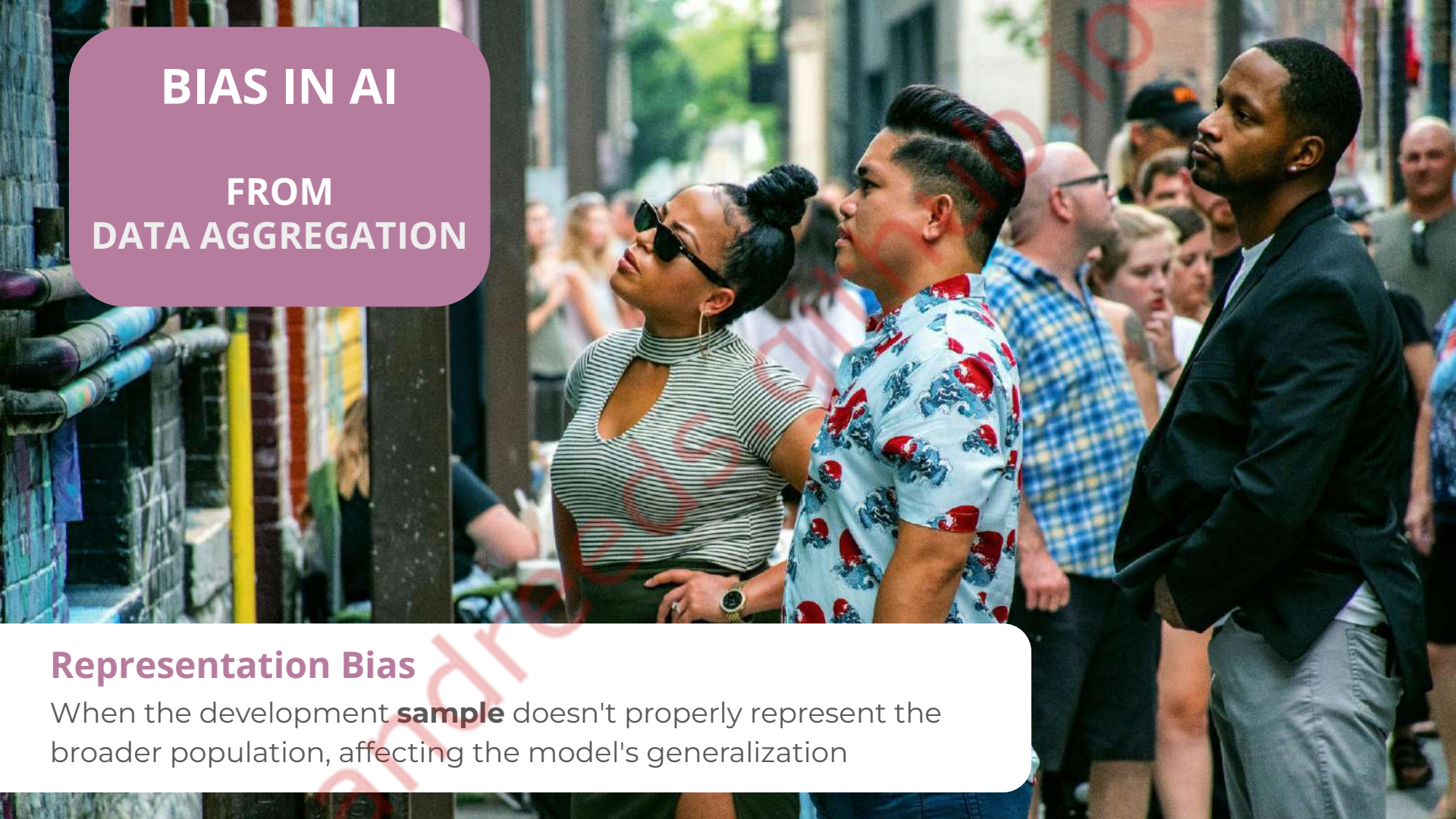
FROM DATA AGGREGATION

Historical Bias

When **data** reflects historical inequalities into model predictions

BIAS IN AI

FROM DATA AGGREGATION



Representation Bias

When the development **sample** doesn't properly represent the broader population, affecting the model's generalization

BIAS IN AI

FROM DATA AGGREGATION

Measurement Bias

When there are flaws in how **data** is **collected** or **measured**, often due to **proxies** that don't accurately capture the desired signals

BIAS IN AI

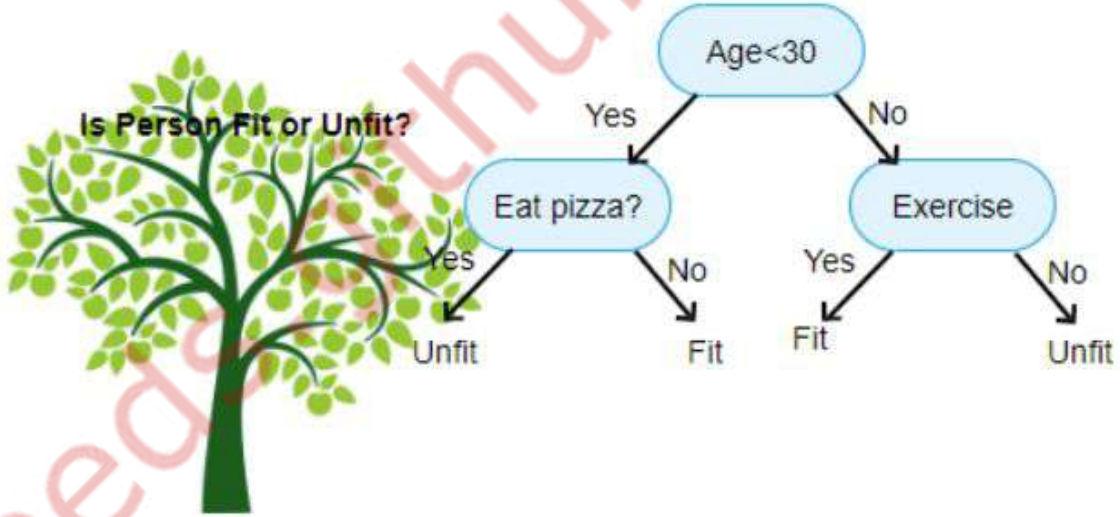
FROM MODEL BUILDING & IMPLEMENTATION

Aggregation Bias

When diverse **groups** are inappropriately **combined** into a single analysis, ignoring meaningful distinctions among them

BIAS IN AI

FROM MODEL BUILDING & IMPLEMENTATION



Learning Bias

When the choice of ML algorithms and their settings may not treat all groups in the data equally

BIAS IN AI

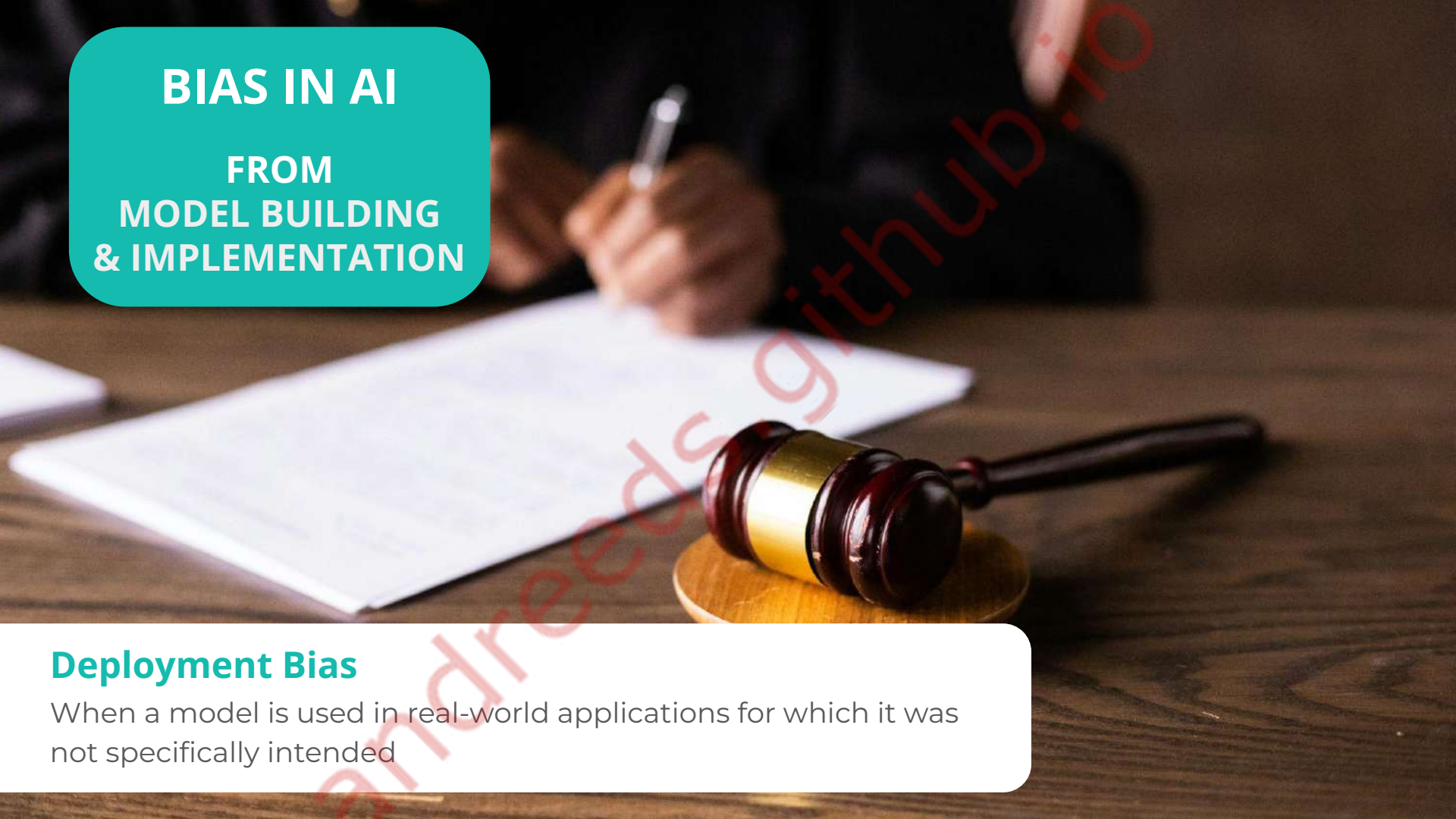
FROM MODEL BUILDING & IMPLEMENTATION

Evaluation Bias

When the data used to evaluate and **benchmark** a model does not represent the actual diversity of the user population

BIAS IN AI

FROM MODEL BUILDING & IMPLEMENTATION



Deployment Bias

When a model is used in real-world applications for which it was not specifically intended

FAIRNESS METRICS

“Everyone agrees that fairness involves treating equal persons equally, and unequal persons unequally, but they do not agree on the standard by which to judge individuals as being equally (or unequally) worthy or deserving.”

—Aristotle

INTRO TO ALGORITHMIC FAIRNESS

Type of Fairness

Within the context of an AI/ML model used for classification

Classification

One type of AI/ML task

Fairness Metrics

Quantitative measures used to assess the fairness of AI models

Pure Math

Understanding the nuances of fairness metrics is essential

BASIC FAIRNESS METRIC TERMINOLOGY

Sensitive attribute

Attribute needing special ethical, legal, or social consideration

Proxy Attribute

Attribute correlated with a sensitive attribute

Parity

Observational measure ensuring metrics are independent of defined groups

Confusion Matrix

Tool measuring accuracy in predictions

CONFUSION METRIC

		True condition			
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive,** Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative,** Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

LINKS

What-If Tool demo - two binary classifiers for predicting salary of over $50k - UCI census income dataset

Datapoint editor | Performance & Fairness | Features

500 datapoints loaded

Configure

Ground Truth Feature: over_50k

WHAT IS GROUND TRUTH?
The feature that your model is trying to predict. More.

Cost Ratio (FP/FN): 1

WHAT IS COST RATIO?
The cost of false positives relative to false negatives. Required for optimization. More.

Slice by: <none>

WHAT DOES SLICING DO?
Shows the model's performance on datapoints grouped by each value of the selected feature.

Fairness

Apply an optimization strategy

Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.

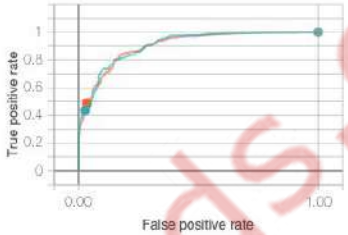
- (•) Custom thresholds
- ( ) Single threshold
- ( ) Demographic parity
- ( ) Equal opportunity
- ( ) Equal accuracy
- ( ) Group thresholds

Explore overall performance

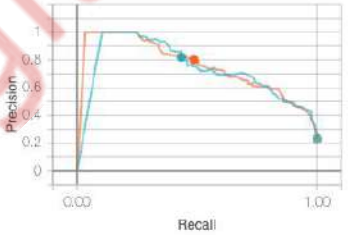
Sort by: Count

Feature Value	Count	Model	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
All datapoints	500	1	0.65	2.2	13.0	84.8	0.57
		2	0.65	2.8	11.8	85.4	0.61

ROC curve (AUCs: 0.90, 0.90)



PR curve (AUCs: 0.71, 0.74)



Confusion Matrix

1	Predicted Yes		Predicted No		Total	
Actual Yes	10.0%	(50)	13.0%	(65)	23.0%	(115)
Actual No	2.2%	(11)	74.8%	(374)	77.0%	(385)
Total	12.2%	(61)	87.8%	(439)		

2	Predicted Yes		Predicted No		Total	
Actual Yes	11.2%	(56)	11.8%	(59)	23.0%	(115)
Actual No	2.8%	(14)	74.2%	(371)	77.0%	(385)
Total	14.0%	(70)	86.0%	(430)		

LINKS

FAIRNESS METRICS EXAMPLES

UNAWARENESS

Ignoring sensitive attributes to achieve fairness in decisions

DEMOGRAPHIC PARITY

Equal decision rates across groups regardless of outcome

EQUAL OPPORTUNITY & EQUALIZED ODDS

Fairness through equal true positive rates and error rates across groups

PREDICTIVE VALUE PARITY

Equal predictive accuracy across different groups

INDIVIDUAL FAIRNESS

Similar treatment for individuals with similar attributes.

Simulating loan decisions for different groups

Drag the black threshold bars left or right to change the cut-offs for loans.
Click on different preset loan strategies.

Loan Strategy

Maximize profit with:

MAX PROFIT

No constraints.

GROUP UNAWARE

Blue and orange thresholds are the same.

DEMOGRAPHIC PARITY

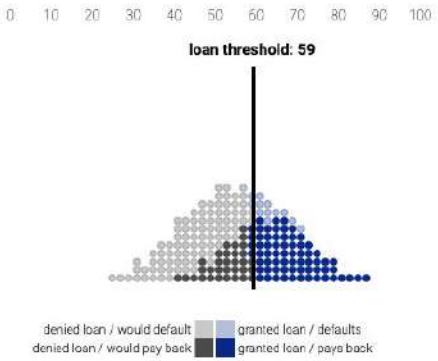
Same fractions blue / orange loans.

EQUAL OPPORTUNITY

Same fractions blue / orange loans to people who can pay them off.

Equal Opportunity
Among people who would pay back a loan, blue and orange groups do equally well. This choice is almost as profitable as demographic parity, and about as many people get loans overall.

Blue Population



Total profit = **30400**

Correct 78%
loans granted to paying applicants and denied to defaulters

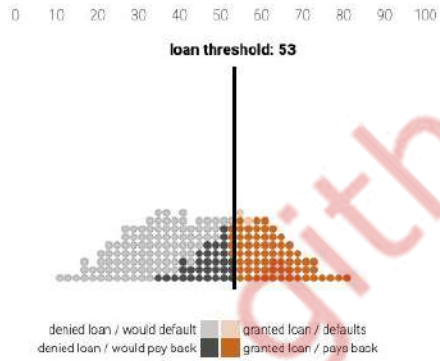
Incorrect 22%
loans denied to paying applicants and granted to defaulters

True Positive Rate 68%
percentage of paying applicants getting loans

Positive Rate 40%
percentage of all applications getting loans

Profit: **11700**

Orange Population



Correct 83%
loans granted to paying applicants and denied to defaulters

Incorrect 17%
loans denied to paying applicants and granted to defaulters

True Positive Rate 68%
percentage of paying applicants getting loans

Positive Rate 35%
percentage of all applications getting loans

Profit: **18700**

LINKS

OPEN SOURCE LIBRARIES

Here are a few common open-source libraries and tools on AI Fairness:

AIF360 [Python/R]

- ❏ IBM Research
- ❏ Last update: 1 month ago
- ❏ Bias mitigation algorithms
- ❏ Fairness metrics
- ❏ Video tutorial

Fairlearn [Python]

- ❏ Microsoft, now community driven
- ❏ Last update: 1 week ago
- ❏ Bias mitigation algorithms
- ❏ Fairness metrics
- ❏ Complete Documentation - Notebook Examples

Aequitas [Python]

- ❏ Carnegie Mellon University
- ❏ Last update: 2 weeks ago
- ❏ Complete Toolkit
- ❏ Complete Documentation - Notebook Examples
- ❏ Aequitas's license does not allow commercial use.

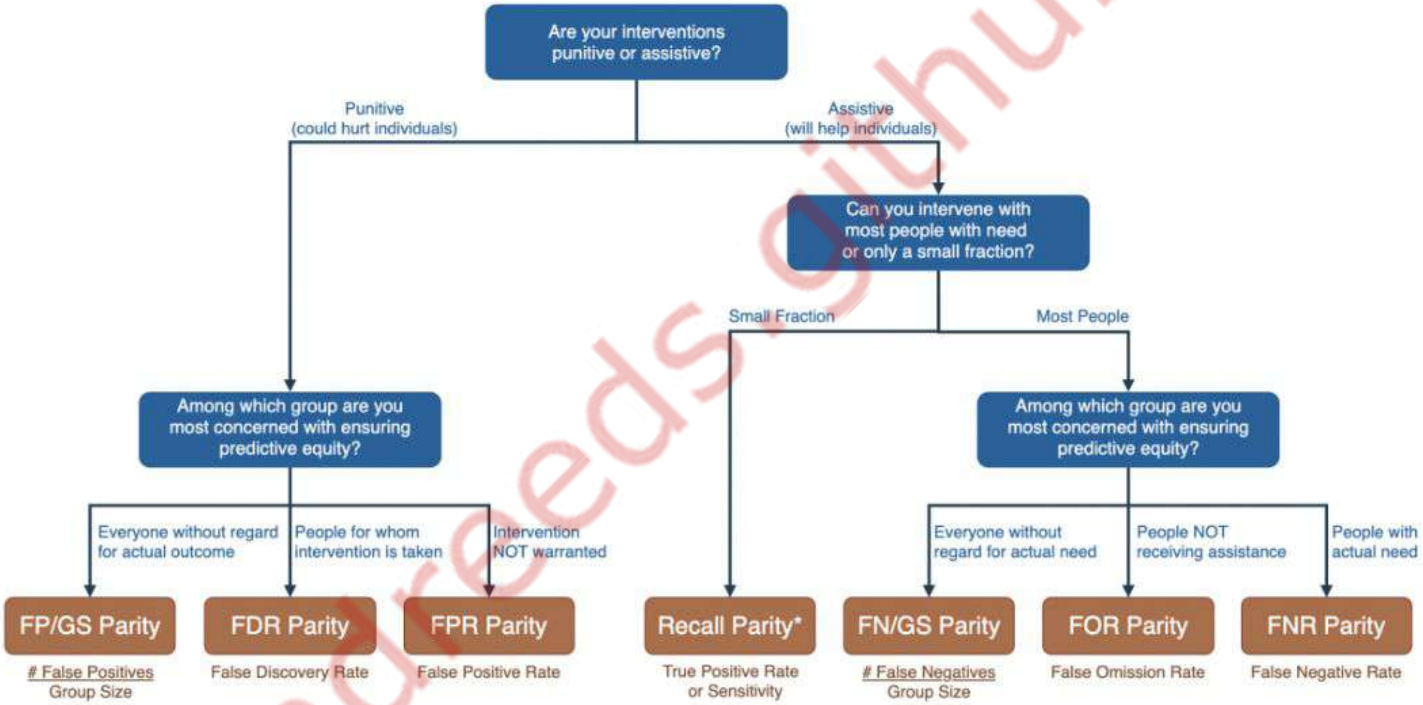
VerifyML [Python]

- ❏ Cylynx
- ❏ Winner Global Veritas Challenge
- ❏ Complete Toolkit
- ❏ Last update: 2 years ago
- ❏ Code Demo

LINKS

Fairness-Short-Tree

FAIRNESS TREE (Zoomed in)



Source: Aequitas

LINKS

Aequitas
Bias & Fairness Audit

Home Code About

Bias and Fairness Audit Toolkit

The Bias Report is powered by Aequitas, an open-source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools.

See an example report on COMPAS risk assessment scores.

Or try out the audit tool using your own data or one of our sample data sets.

Get Started!

LINKS

[X]ai

PART 2

Responsible AI

LINKS

RESPONSIBLE AI DEFINITIONS

Whats is the difference between **Understandability**, **Interpretability**, **Comprehensibility**, **Explainability**, and **Transparency**?

RESPONSIBLE AI KEY CONCEPTS

UNDERSTANDABILITY

What?

COMPREHENSIBILITY

How?

INTERPRETABILITY

Why? When?
Who? Which?

EXPLAINABILITY

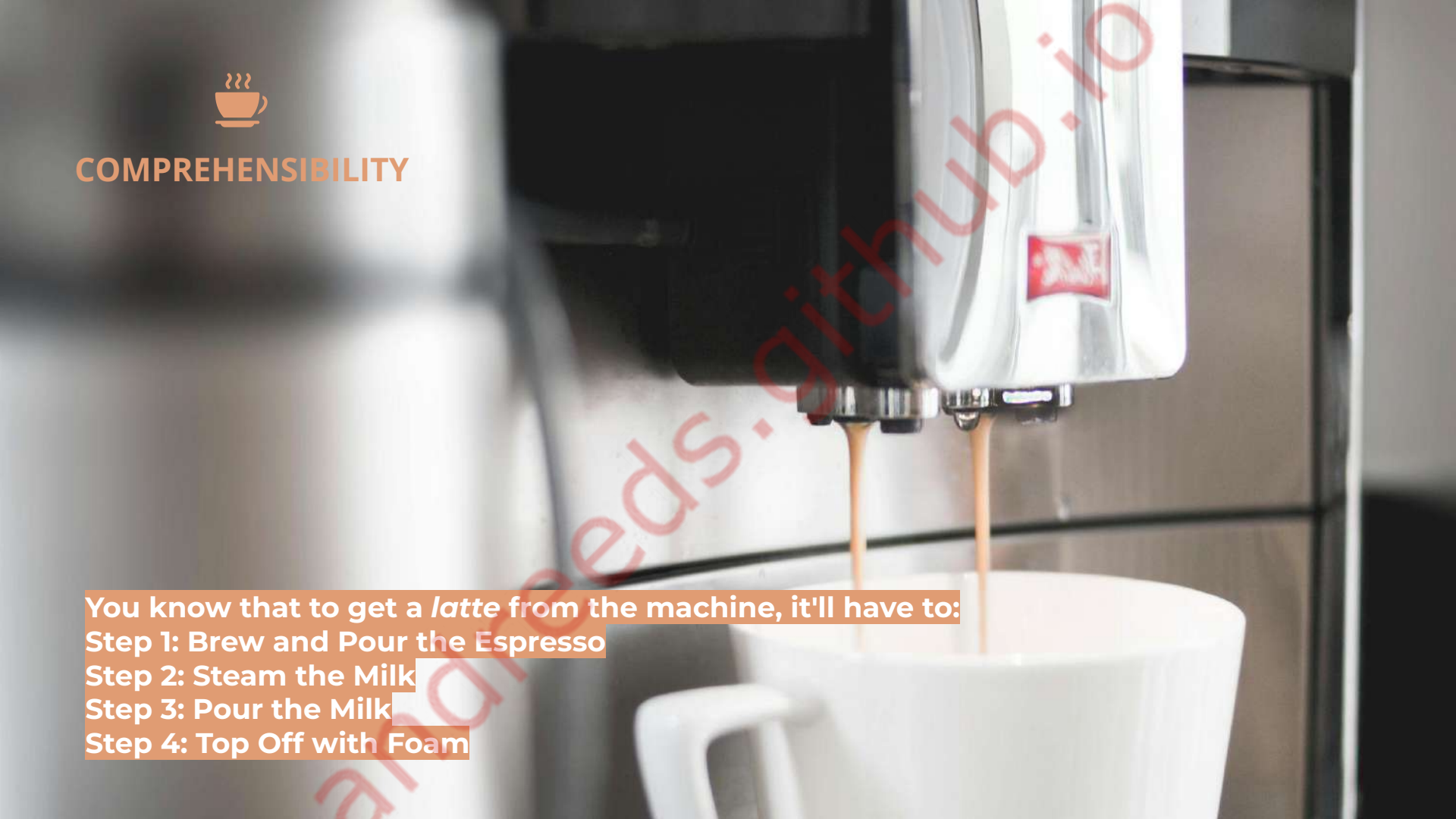
Why? What?
Who? When?
How? Which?

TRANSPARENCY

No questions needed.

UNDERSTANDABILITY

You know to get a *latte* from the machine



COMPREHENSIBILITY

You know that to get a *latte* from the machine, it'll have to:
Step 1: Brew and Pour the Espresso
Step 2: Steam the Milk
Step 3: Pour the Milk
Step 4: Top Off with Foam

INTERPRETABILITY

You know that if your *latte* is not great, it's because the machine got the milk too hot - 40 degrees for small drinks up to 155 for big ones. Likely they do not account for the lag with the thermometer.

EXPLAINABILITY (XAI)

You know that to get a *latte* from the machine, you see it
Step 1: Brew and pour 2 shots of espresso.
Step 2: Steam 1/2 cup milk to 150 F
Step 3: Pour the steamed milk over the espresso, using a spoon to hold back the foam.
Step 4: Top off the drink with the reserved foam

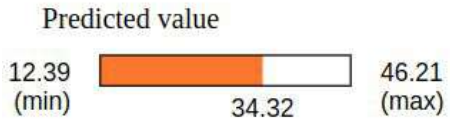

TRANSPARENCY
You ask the barista for a *latte*.

XAI METHODS

LIME

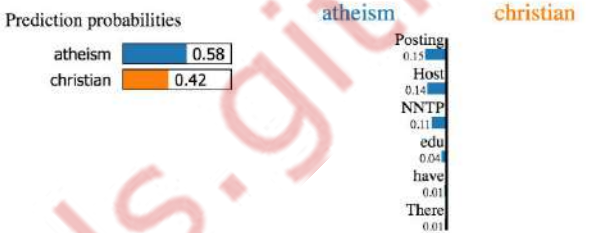
LINKS

Tabular

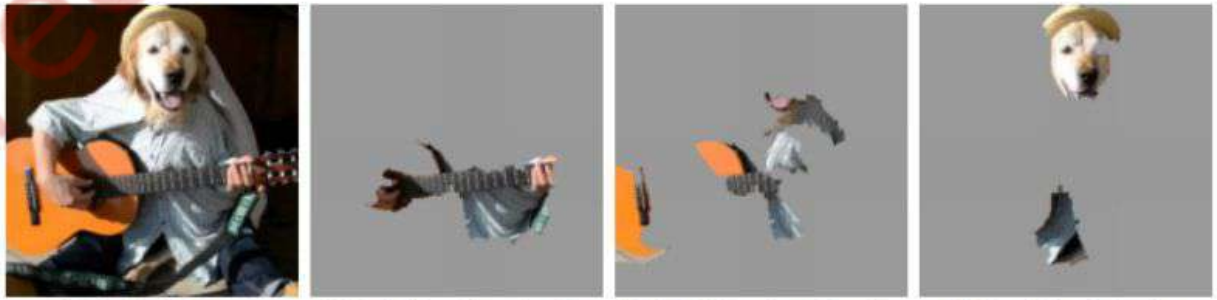


Feature	Value
RM	7.27
LSTAT	6.05
TAX	254.00
INDUS	6.41
PTRATIO	17.60

Text



Image



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

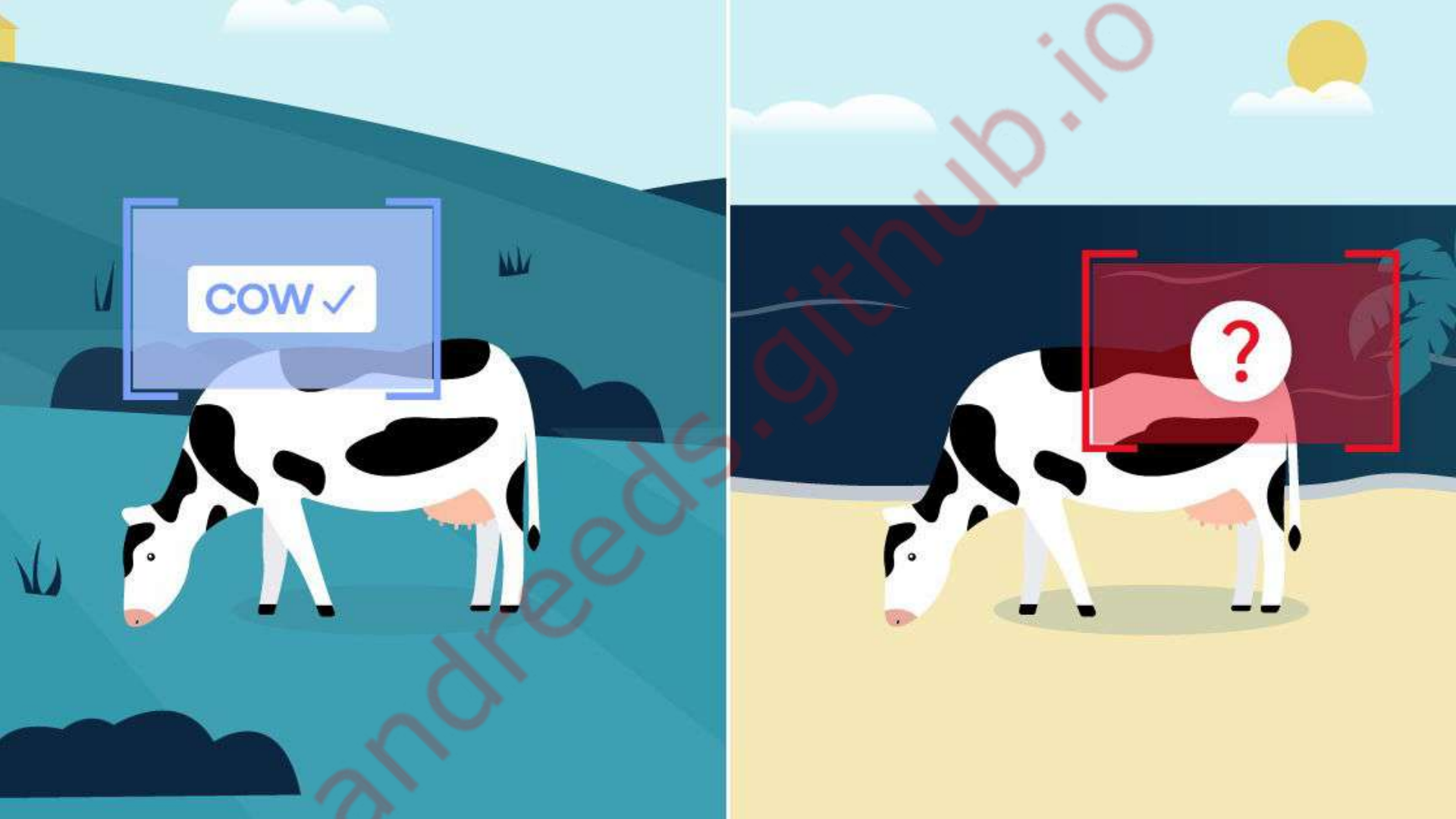
SHAP

LINKS

TRANSPARENT AI

VERSUS

DL + XAI


COW ✓
?


[X]ai
COFFEE BREAK

PART 3

AI Regulations, Standards & Guidelines

LINKS

REGULATIONS

United Kingom

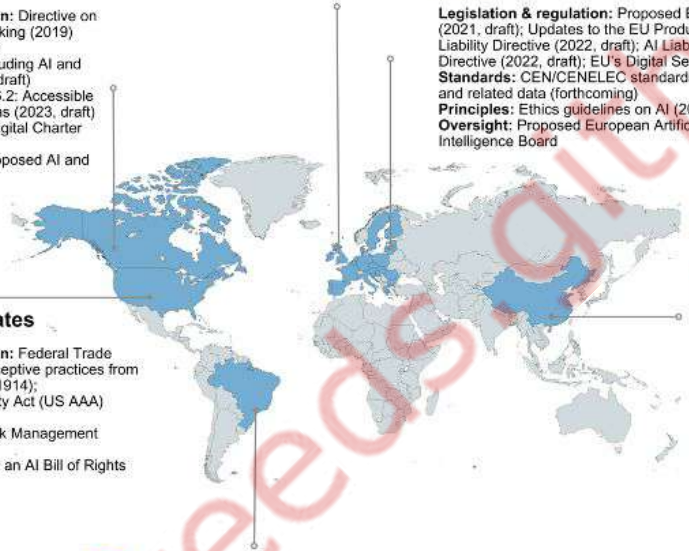
Legislation & regulation: Online Safety Bill (2022, draft); Data Protection and Digital Information Bill (2023, draft)
Standards: Algorithmic Transparency Standard (Central Digital Data Office, 2021)
Principles: A pro-innovation approach to AI regulation (2023)

Canada

Legislation & regulation: Directive on Automated Decision-Making (2019) Bill C-27; Digital Charter Implementation Act, including AI and Data Act (AIDA) (2022, draft)
Standards: CAN-ASC-6.2: Accessible and Equitable AI Systems (2023, draft)
Principles: Canada's Digital Charter (2019)
Oversight: Minister; Proposed AI and Data Commissioner

European Union

Legislation & regulation: Proposed EU AI Act (2021, draft); Updates to the EU Product Liability Directive (2022, draft); AI Liability Directive (2022, draft); EU's Digital Services Act
Standards: CEN/CENELEC standards for AI and related data (forthcoming)
Principles: Ethics guidelines on AI (2018)
Oversight: Proposed European Artificial Intelligence Board



United States

Legislation & regulation: Federal Trade Commission Act, for deceptive practices from deepfakes or chatbots (1914); Algorithmic Accountability Act (US AAA) (2022, draft)
Standards: NIST AI Risk Management Framework (2023)
Principles: Blueprint for an AI Bill of Rights (2023

China

Legislation & regulation: Chinese Internet Information Service Algorithmic Recommendation Provisions (2021); Opinion on Strengthening the Ethics and Governance of Science and Technology (2022)
Standards: National Standards for Autonomous Vehicle Testing (2018)
Principles: New Generation AI Ethics Specifications (2019); New Generation AI Code of Ethics (2021); White Paper on Trustworthy AI (2021); Internet Information Service Algorithmic Recommendation Management Provisions (2021)

Brazil

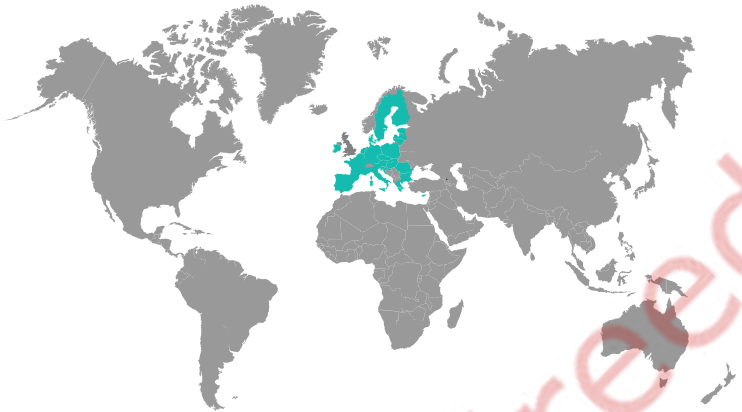
Legislation & regulation: Report and proposed substitute text for draft bills 5051/2019, 21/2020 and 872/2021 (2022, draft); Bill 705 on the compatibility of AI use in the public sector with ESG practices (2022, draft)
Standards: Incorporation of international standards National standards by the Brazilian Association of Technical Norms (ABNT)
Principles: Art. 3 of the proposed substitute text for draft bills 5051/2019, 21/2020 and 872/2021 (2022, draft)

Intergovernmental Organisations

Legislation & regulation: Council of Europe Convention on AI, Human Rights, Democracy and the Rule of Law (2023, draft)
Standards: ISO 31000 Risk management (2009, 2018); ISO/IEC 23053:2022 Framework for AI Systems Using Machine Learning (ML) (2022)
Principles: OECD Recommendation of the Council on AI (2019); UNESCO Recommendation on the Ethics of AI (2021)

LINKS

EU



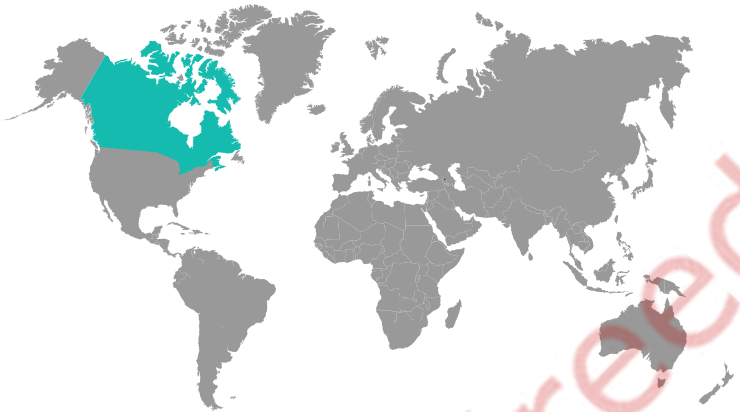
AI ACT

AI systems are broadly defined, with a focus on autonomy.

Key takeaways

- Risk-Based Approach
- Banned Practices
- Transparency Obligations
- Market Surveillance

CANADA



BILL C-27

Bill comprised of three acts, one is the AI & Data Act (AIDA)

- Committee Stage
- Government institutions excluded
- Two purposes: regulate trade and prohibit harm
- Prevention of biased **outputs**
- High impact system not clear
- Minister of Innovation Powers

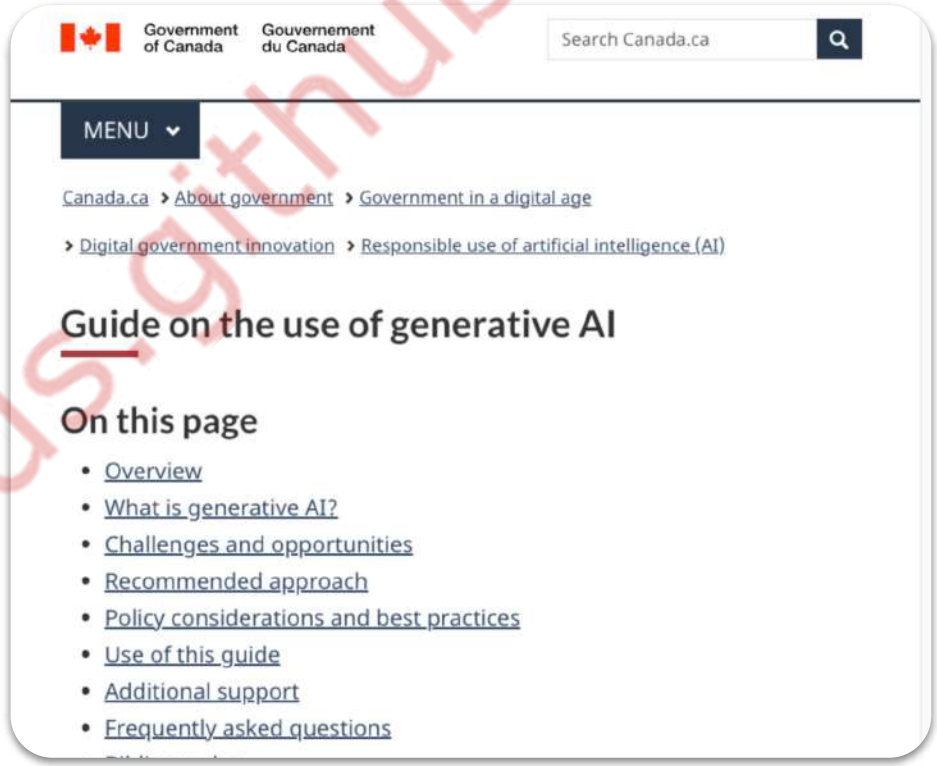
GUIDELINES

EU AI ACT COMPLIANCE CHECKER

LINKS

GOVERNMENT OF CANADA'S

GUIDE ON THE USE OF GENERATIVE AI



LINKS

GOVERNMENT OF CANADA'S

ALGORITHMIC IMPACT ASSESSMENT TOOL

LINKS

OECD publishing

OECD FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS

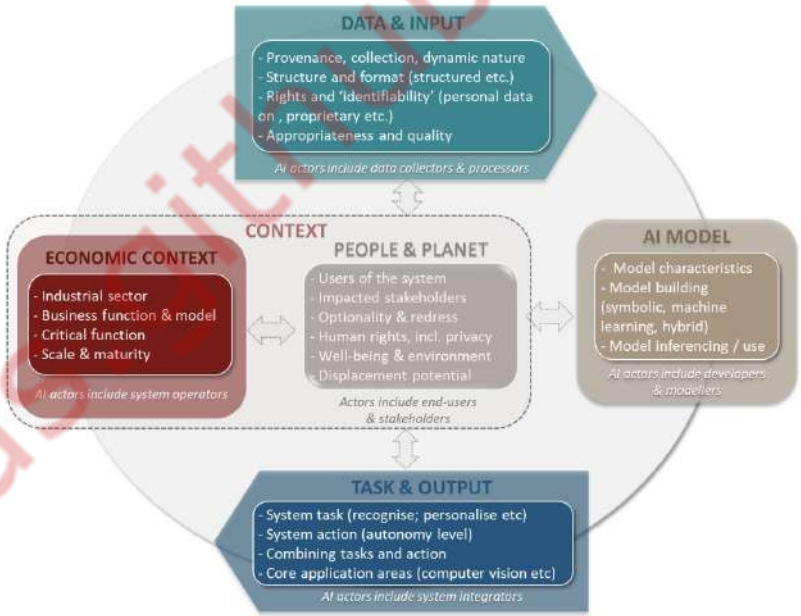
OECD DIGITAL ECONOMY PAPERS

February 2022 **No. 323**

Federal Ministry of Labour and Social Affairs

OECD
BETTER POLICIES FOR BETTER LIVES

Figure 2. Characteristics per classification dimension and key actor(s) involved



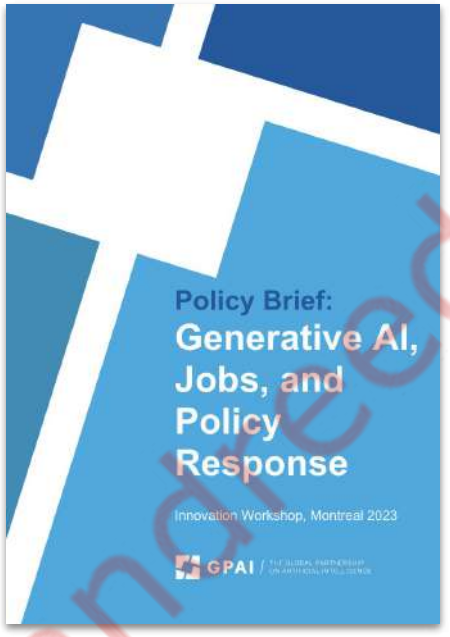
Note: Actors are illustrative, non-exhaustive and notably relevant to accountability.
Source: Based on the work of ONE AI and the AI system lifecycle work of AIGO (OECD, 2019f[2]).

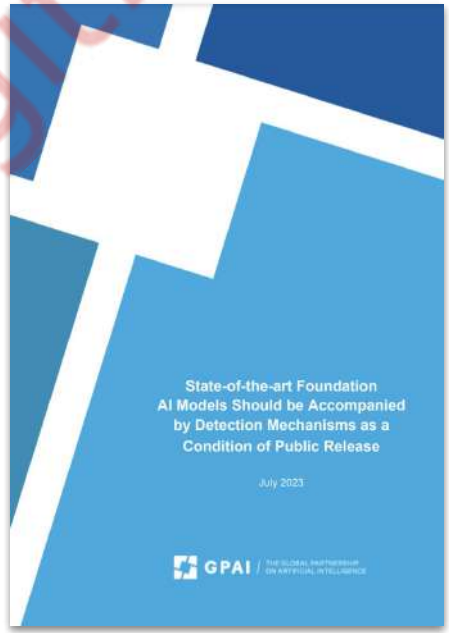
LINKS

Follow us

GPAI / THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

HOME ABOUT COMMUNITY OUR WORK EVENTS





STANDARDS

IEEE

ETHICALLY ALIGNED DESIGN

LINKS

Digital Governance
Standards Institute

ETHICAL DESIGN AND USE OF AUTOMATED DECISION SYSTEMS

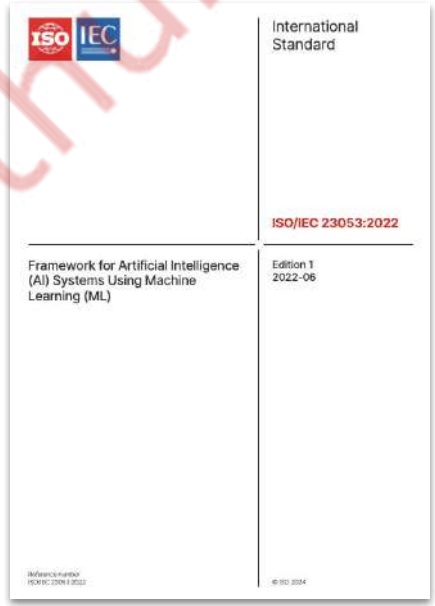
LINKS

[X]ai

[X]ai

ISO/IEC 23053:2022

FRAMEWORK FOR AI SYSTEMS USING ML



LINKS

CONCLUSION

PART 1

- Bias
- Fairness

PART 2

- RAI Definitions
- XAI

PART 3

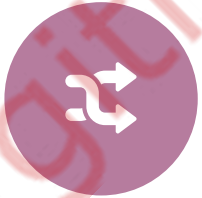
- Regulations
- Standards
- Guidelines

LINKS

MORE ESSENTIALS

PRIVACY

Adversarial attacks



CAUSALITY

Correlation is not necessarily causation

AI & PEOPLE

How AI affects and is perceived by people

FRAMEWORKS

Ethics charter

[X]ai

THE **TRUTH** ABOUT AI

Separating **Fact** from **Fiction**

by Dr. André dos **Santos**

Computer Science > Machine Learning

[Submitted on 6 Apr 2023 (v1), last revised 29 Oct 2023 (this version, v3)]

Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models

Pengfei Li, Jianyi Yang, Mohammad A. Islam, Shaolei Ren

The growing carbon footprint of artificial intelligence (AI) models, especially large ones such as GPT-3, has been undergoing public scrutiny. Unfortunately, however, the equally important and enormous water (withdrawal and consumption) footprint of AI models has remained under the radar. For example, training GPT-3 in Microsoft's state-of-the-art U.S. data centers can directly evaporate 700,000 liters of clean freshwater, but such information has been kept a secret. More critically, the global AI demand may be accountable for 4.2 -- 6.6 billion cubic meters of water withdrawal in 2027, which is more than the total annual water withdrawal of 4 -- 6 Denmark or ha... the most pressing challenges share... aging water infrastructures. To resp... and lead by example by addressing... water footprint of AI models, and a... Finally, we highlight the necessity o... AI.

Comments: New updates include discussion... footprint based on Microsoft's ...

Subjects: Machine Learning (cs.LG); Arti...

Cite as: arXiv:2304.03271 [cs.LG]

(or arXiv:2304.03271v3 [cs.LG]

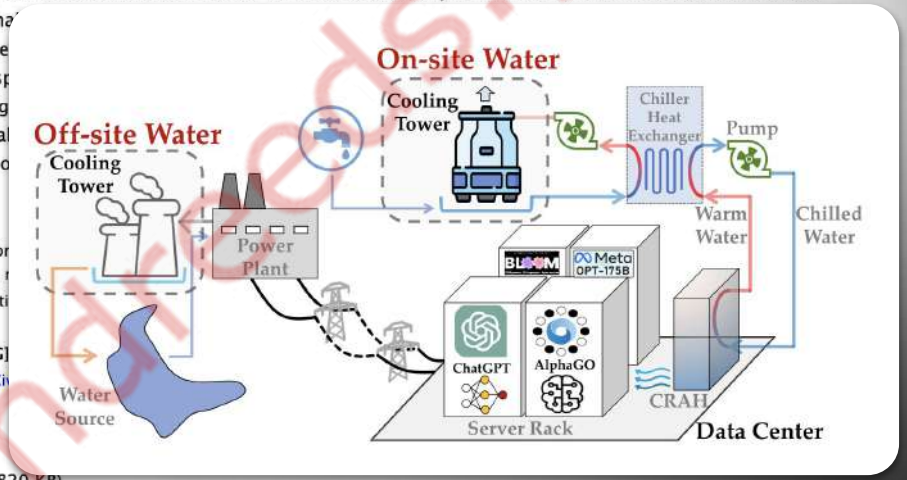
https://doi.org/10.48550/arXi...

Submission history

From: Shaolei Ren [view email]

[v1] Thu, 6 Apr 2023 17:55:27 UTC (2,820 KB)

[v2] Wed, 25 Oct 2023 07:56:21 UTC (232 KB)



Export BibTeX Citation

Bookmark


BE HUMAN.

[X] ai plain

PROFESSIONAL DEVELOPMENT

AI EXPLAINED

AI SAFETY

AI LEADERSHIP

EXPERTS NETWORKING

AI CONSULTING

by Dr. André dos **Santos**

[X]ai