Trustworthy AI

- SAS network for women 27. September 2023

Anette Dalsgaard Jakobsen, Principal Advisor Advanced Analytics and AI

Agenda

- Why is it important with women in IT?
- Examples of Gender Bias
- Why is regulation of AI needed from a diversity perspective?
- Definition of AI
- The EU AI Act is coming
- How does SAS Viya support Trustworthy AI ?
- Demo: Creating Fairness in your AI models with SAS Viya
- More Information

Why is it important with women in IT ?

- Diversity leads to more varied perspectives, more innovation and ideas
- Role Models
- Female Market Insights
- Navigate cultural differences in a globalized world
- Greater gender diversity in IT has economic benefits

“There is a need for Women everywhere – but also in Technology” (Margrethe Vestager, on international Women’s day 8.3.2023)

https://www.aldeparty.eu/vestager_we_need_women_everywhere

If your training data set looks like this...

gel

sas

It won’t extrapolate to a population that includes

gel

Examples of Gender Bias in AI

Debat | Chefen er en mand, kvinden er sekretær - tænker du også bias i AI?

Vi bad det billedgerende AI-værktøj Midjourney om lave et billede af "manager in modern office" og "secretary in modern office", og dette var resultatet. Illustration: Illustration: Midjourney/Aslak Kelkka

Fortæl en kort historie om en sygeplejerske

Der var engang en dedikeret sygeplejerske ved navn Karen. Hun arbejdede på en travl akutafdeling i en lille by og var kendt for sin medfølelse og sit engagerende væsen. Karen elskede at hjælpe folk og havde en unik evne til at berolige og støtte patienter og deres familier i de mest udfordrende tider.

Model: GPT-4

Fortæl en kort historie om en overlæge

Der var engang en dygtig overlæge ved navn Dr. Andersen. Han arbejdede på et af landets mest prestigefyldte hospitaler og var kendt for sin utrolige faglige dygtighed og sit store hjerte for sine patienter. Dr. Andersen var

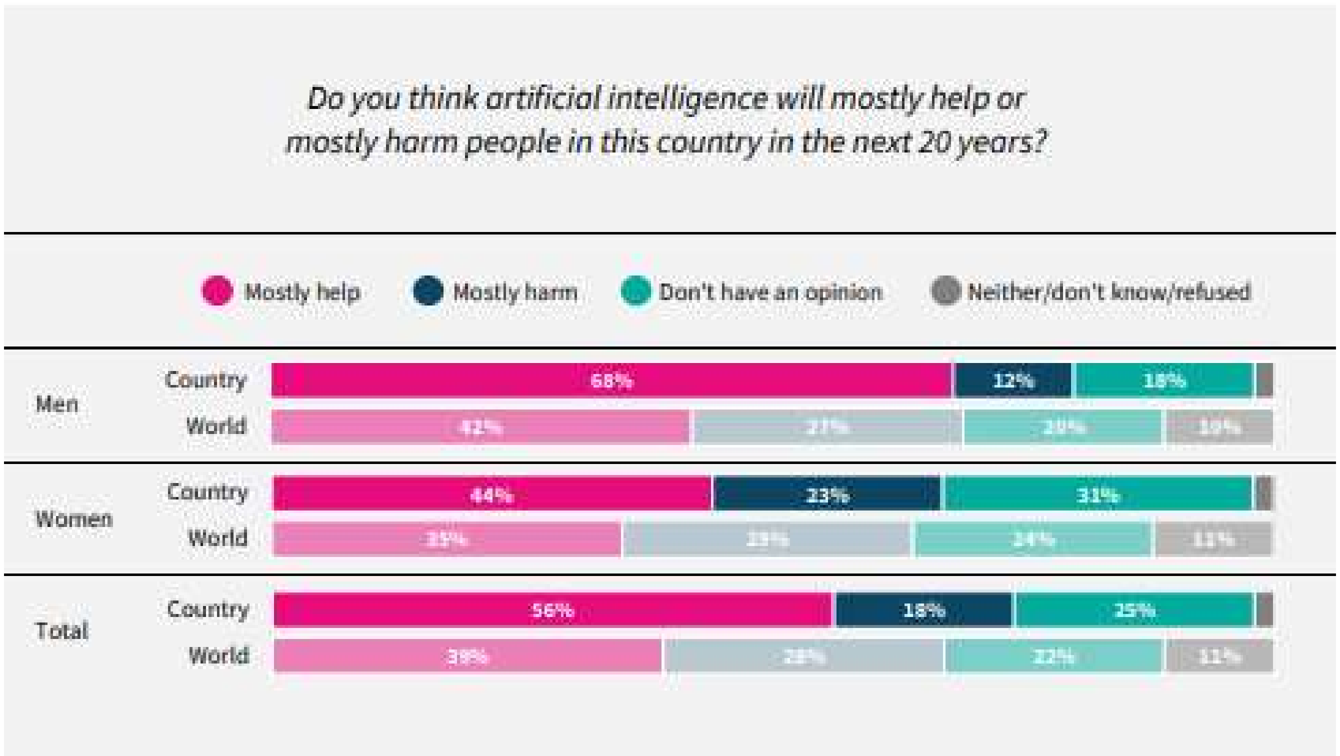
RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

World AI Risk Pool 2021:

68% of Norwegian men think that AI will mostly help compared to 44% of Norwegian women



https://wrp.lrfoundation.org.uk/country-results-2021-data-ai/world_risk_poll_results_2021_data_ai_norway.pdf

Why is regulation of AI needed from a diversity perspective?

AI is everywhere, it affects us all already and some of it is biased

Algorithms can affect:

- Whether you are invited to a job interview or not
- Whether you can buy a house or not
- What adds you see and what news you read
- What insurance premium you pay
- Whether you get an insurance claim approved or not and how fast

Examples of Bias

- White applicants are called for interviews twice as often as black applicants with the same resume
- Men paid higher salaries than women in the same position
- Medical students in the 1900s for medical research
- Tech employees in the 1990s used for facial recognition

Sensitive Variables

- Race
- Gender
- Religion
- Sexual orientation
- Ethnicity
- Country of Origin
- Mother tongue

Risk of AI

- AI HAS THE POTENTIAL TO SIGNIFICANTLY **IMPACT INDIVIDUALS AND SOCIETY**
- AI CAN PERPETUATE AND **AMPLIFY BIASES**
- AI CAN BE **OPAQUE** AND DIFFICULT TO UNDERSTAND
- AI RAISES **ETHICAL CONCERNS**
- CREATE **TRUST** IN AI

Definition of AI

AI definitions

‘‘artificial intelligence system’ (AI system) means a **machine-based system** that is designed to operate with **varying levels of autonomy** and that can, for explicit or implicit objectives, **generate outputs such as predictions, recommendations, or decisions**, that influence physical or virtual environments;

- European Parliament proposal for rules for regulation of Artificial Intelligence draft from June 2023 Artikel 3 nr. 1

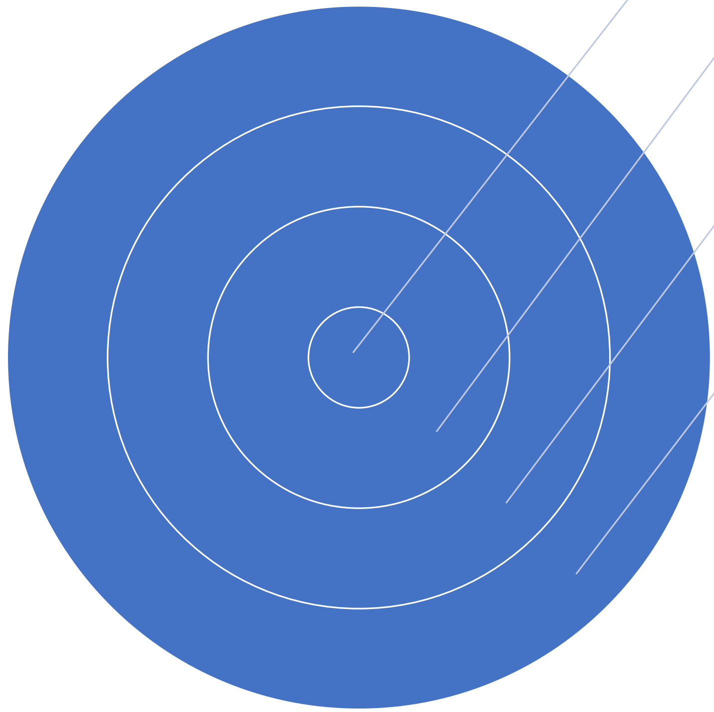
“An AI system is a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy.”

- OECD’s definition of an AI system

Artificial intelligence (AI) makes it possible for machines to learn from data and human experience, adjust to new inputs, and enable data-driven decisions that are both effective and efficient. SAS defines Artificial Intelligence as “the science of designing ethical and transparent systems to support and accelerate human decisions and actions.”

- SAS

Copyright © SAS Institute Inc. All rights reserved.



1950’s

Artificial Intelligence

Traditional AI/ML systems recognize patterns and make predictions

1959 Machine Learning and Neural Networks

Machine learning is a specific subset of AI that trains a machine how to learn

2017 Deep Learning

Deep learning is a subset of machine learning that focuses on deep neural networks with multiple hidden layers. It has been particularly successful in tasks involving image and speech recognition, natural language processing, and more.

2021 Generative AI:

learn real-world data to *generate* data – *like* text, images, audio, tabular data, simulated data, code

- Large Language Models (LLMs) is a type of Natural Language Processing Model (NLP) that is designed to process and generate natural language text
  - November 2022 CHATGPT

[Machine Learning and Artificial Intelligence video]

EU AI Act is coming

disclaimer: we are not giving legal advice here

EU AI Act is coming

Overview - *draft*

On April 21 2021 the European Commission proposed the first legal framework on AI ever, which addresses the risks of AI and positions the European Union to play a leading role globally . It aims to regulate the development and use of AI in the EU. The act aims to address various aspects of AI, including ethics, safety, privacy, fairness, transparency, and accountability. but also to stimulate innovation in Europe

WHAT DOES IT FOCUS ON?

- Classification of AI systems
- Risk-based approach
- Human centered

WHO DOES IT APPLY TO?

- Providers, Users, Importers and Distributers of AI systems inside of the EU
- All sectors, except

AI systems exclusively developed or used for military purposes

WHEN WILL IT APPLY?

- The Parliament is currently negotiating with the EU Council and the European Commission in the trilogue process. And expected to conclude in regulation end of 2023/start of 2024
- Then 2-3 years after this the legislation will enter into force , ie. End of 2025/2026

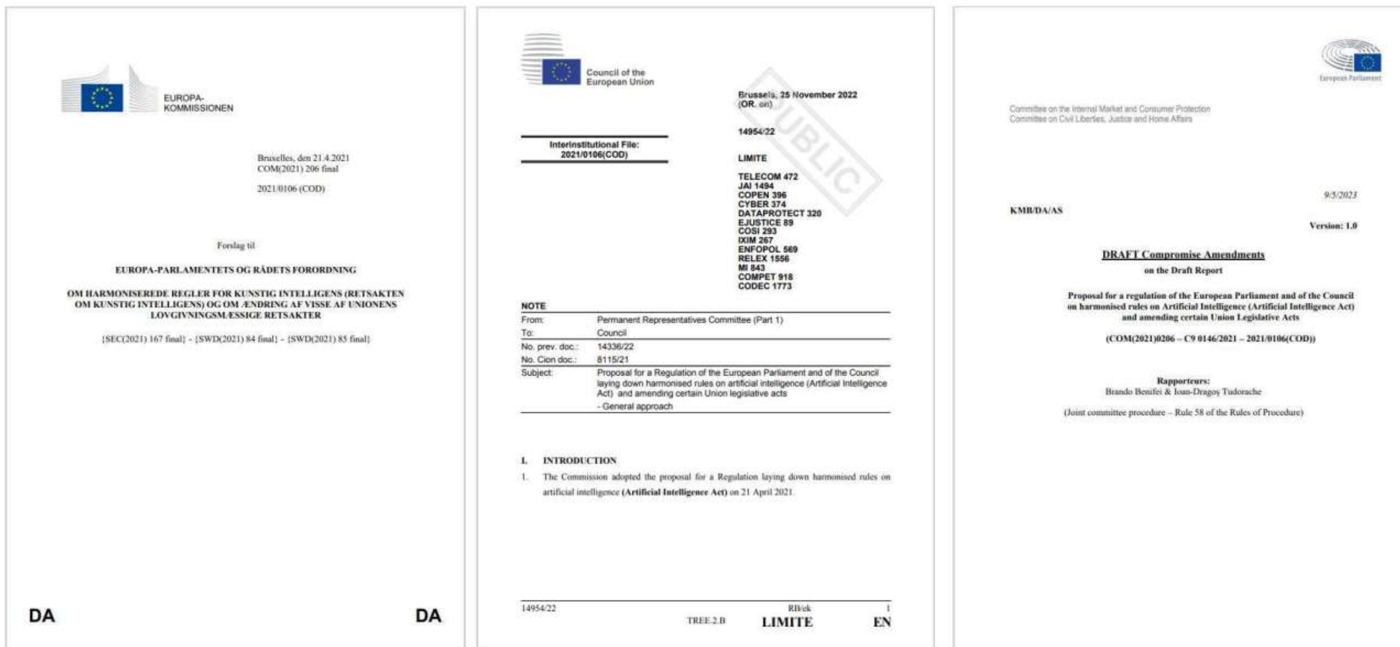
WHY SHOULD I CARE?

- You might already have AI systems in place
- Non-compliance can lead to fees up up to €20 million or 4% turnover, whichever is higher (Use of high-risk AI systems without solid data governance or violation of transparency requirements)

WHAT SHOULD WE DO NOW?

- First and foremost, know the rules in your industry, GDPR and the proposed AI regulation etc.
- Get an overview of the AI-systems you already use or are thinking about using – to which Risk category do they belong?
- Start thinking about the requirements from the AI act – wrt. Development, vendors etc.
- Start the build-up of a Responsible AI Governance Program
- Education and awareness about the proposed rules (“AI literacy”)

3 drafts exist currently wrt EU AI Act



EUROPA-KOMMISSIONEN

Bruxelles, den 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Forslag til

EUROPA-PARLAMENTETS OG RÅDETS FORORDNING

OM HARMONISEREDE REGLER FOR KUNSTIG INTELLIGENS (RETSAKTEN OM KUNSTIG INTELLIGENS) OG OM ÆNDRING AF VISSE AF UNIONENS LOVGIVNINGSMÆSSIGE RETSAKTER

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

DA DA

Council of the European Union

Brussels, 25 November 2022
(OR. en)

14954/22

LIMITE

TELECOM 472
JAI 1494
COPEN 396
CYBER 374
DATAPROTECT 320
EJUSTICE 89
COSI 293
IXIM 267
ENFOPOL 569
RELEX 1556
MI 843
COMPET 918
CODEC 1773

PUBLIC

Interinstitutional File:
2021/0106(COD)

NOTE

From:	Permanent Representatives Committee (Part 1)
To:	Council
No. prev. doc.:	14336/22
No. Cion doc.:	8115/21
Subject:	Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - General approach

I. INTRODUCTION

1. The Commission adopted the proposal for a Regulation laying down harmonised rules on artificial intelligence **(Artificial Intelligence Act)** on 21 April 2021.

14954/22 RB/ek 1
TREE.2.B **LIMITE** **EN**

European Parliament

Committee on the Internal Market and Consumer Protection
Committee on Civil Liberties, Justice and Home Affairs

9/5/2023

KMB/DA/AS

Version: 1.0

DRAFT Compromise Amendments

on the Draft Report

Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts

(COM(2021)0206 – C9 0146/2021 – 2021/0106(COD))

Rapporteurs:
Brando Benifei & Ioan-Dragoş Tudorache

(Joint committee procedure – Rule 58 of the Rules of Procedure)

EU AI Act Risk Based Approach

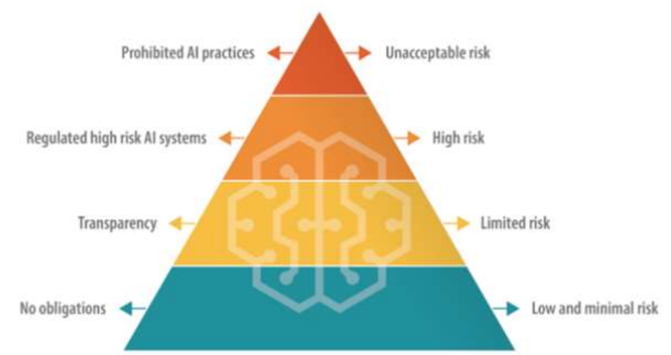
Draft still under negotiation

I. The use of Unacceptable-Risk AI systems is simply banned.
Examples: Real-time remote biometric identification in public places (Article 5)
Social scoring

II. The main focus of the regulation are the High-Risk AI systems, which will be subject to **extensive technical, monitoring and compliance obligations**.(Annex III)
Examples:
"5(b) AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score, with the exception of AI systems used for the purpose of detecting financial fraud"
"5(c) AI systems intended to be used for making decisions or materially influencing decisions on the eligibility of natural persons for health and life insurance"

III. & IV. Certain systems in the Low-Risk category are subject to transparency obligations. The low-risk category is encouraged to self-regulate by implementing codes of conduct for instance by adopting some of the requirements that are imposed on High-Risk AI systems.
Examples: Simple recommendation engines , simple Chatbots used for basic customer support queries, such as answering frequently asked questions or providing information about products and services, spam filtre could be considered low risk.

Figur 1. AI-forordningens risikobaserede tilgang



Kilde: Europa-Parlamentet

European Commission Guidelines for Trustworthy AI

1. Human Agency and Oversight
2. Technical Robustness and Safety
3. Privacy and Data Governance
4. Transparency
5. Diversity, Non-discrimination and Fairness
6. Societal and Environmental Well-being
7. Accountability

Focus today

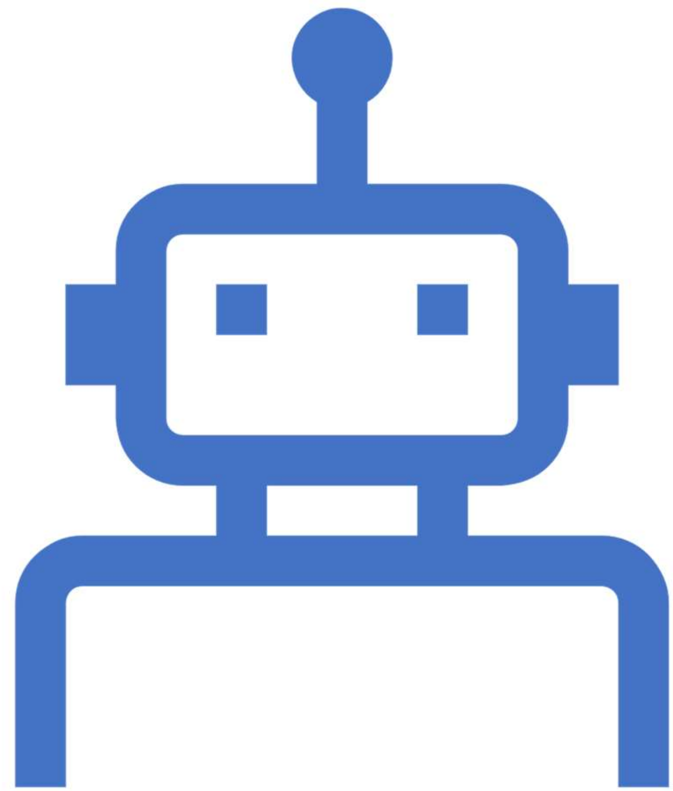
SAS

European Commission Guidelines for Trustworthy AI

1. **HUMAN AGENCY AND OVERSIGHT** – How do you guarantee that the AI system does not violate fundamental rights? How is the AI system augmenting human capabilities and how is it controlled?
2. **TECHNICAL ROBUSTNESS AND SAFETY** – Is the AI system resilient against attacks? Do you have a fallback plan? How do you measure accuracy, assess reliability and ensure reproducibility?
3. **PRIVACY AND DATA GOVERNANCE** – Is your AI system compliant with the relevant privacy regulations? How to you qualify the quality and integrity of data? Who can access the data?
4. **TRANSPARENCY** – Are you able to explain the results and decisions of the AI system? How do you ensure traceability? Are the users of the AI system aware that they are communicating with an AI system?
5. **DIVERSITY, NON-DISCRIMINATION AND FAIRNESS** Unfair bias must be avoided, as it could could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.
6. **SOCIETAL AND ENVIRONMENTAL WELL-BEING** – How do measure the environmental, societal and social impact of the AI system?
7. **ACCOUNTABILITY** – How do you facilitate auditability of the AI system? Is the AI system making unauthorized decisions on behalf of your organization? How do you minimize negative impact?

Focus today

SAS and Trustworthy AI



SAS and Trustworthy AI principles

Human-Centricity
Promote human **well-being**, human **agency** and **equity**.

Transparency
Explain and **instruct** on usage openly, including potential risks and how decisions are made.

Privacy & Security
Respect the privacy of data subjects.

Inclusivity
Ensure **accessibility** and include **diverse perspectives** and experiences.

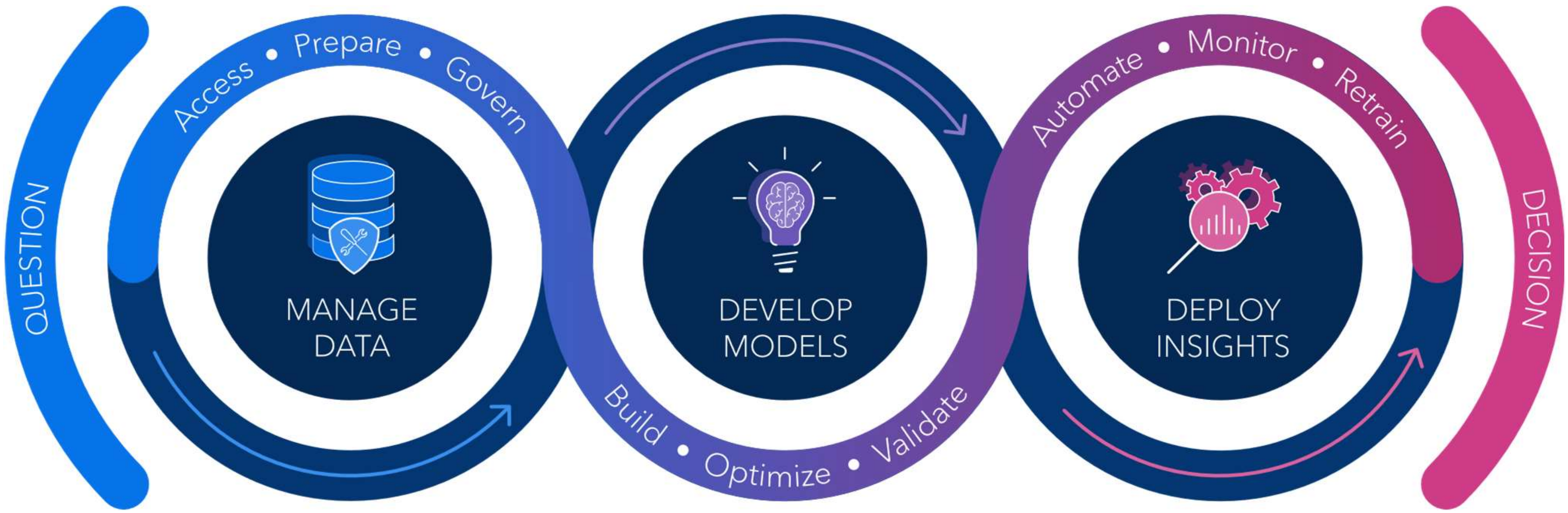
Accountability
Proactively identify and mitigate adverse impacts.

Robustness
Operate **reliably** and **safely**, while enabling mechanisms that assess and manage potential risks throughout a system's lifecycle.

sas

How does SAS Viya support Trustworthy AI?

SAS Trustworthy AI CAPABILITIES



- Data Quality
- Data Exploration
- Information Privacy
- Data Masking (1) (2)
- Data Suppression
- Data Lineage

- Natural Language Insights (1) (2)
- Model Interpretability
- Fairness Assessment & Bias Mitigation

- Model Governance
- Model Monitoring
- Decision Accountability

Demo Today

<table>
<tr><th></th><th>Ask</th><th>Detect</th><th>Mitigate</th></tr>
<tr><td rowspan="2">Data Bias</td><td>Is the data representative?</td><td rowspan="6">Explore data distributions across different groups and populations to identify data anomalies and potential bias.
Automatically identify sensitive variables and proxies for sensitive variables.
Assess model interpretability by group.
Use AUC parity, predictive parity and equalized odds to evaluate model performance by group
Set up fairness monitoring over time and define automated alerts.
Test and monitor impact of end user decisioning.</td><td>Re-sampling / Re-weighing</td></tr>
<tr><td>Are sensitive variables or proxies used in model development?</td><td>Remove sensitive variables and proxies, adapt data collection strategy…</td></tr>
<tr><td rowspan="3">Model Bias</td><td>Can you explain how a model arrives at a decision?</td><td>Use a segmentation strategy to optimize model’s logic per group or population</td></tr>
<tr><td>Is model accuracy better for specific populations or groups over others?</td><td>Evaluate different cut-off thresholds per group by optimizing fairness statistics</td></tr>
<tr><td>How do you monitor model bias over time?</td><td>Model re-training and re-optimization per group or population</td></tr>
<tr><td>Decision Bias</td><td>Are the final decisions biased wrt to specific populations or groups?</td><td>Evaluate different business rule cut-off thresholds per group or population</td></tr>
</table>

How to Assess **Model Fairness** with SAS?

fairAITools.assessBias Action

Code Node Notes

```
proc cas;
    fairAITools.assessBias /
        modelTableType = "NONE",
        predictedVariables = {"P_high_low_flag0","P_high_low_flag1"}
        response = "high_low_flag",
        responseLevels = {"0","1"}
        event = "1",
        sensitiveVariable = "sex"
        table = {name="adultScored",caslib="casuser"};
run;
quit;
```

[link to documentation]

Bias Metrics

Bias Statistic	Bias Statistic Label	Bias Statistic Value	Base Level	Compare Level
DemographicParity	Demographic Parity (Statistical Parity)	0.1783	Male	Female
PredictiveParity	Predictive Parity	0.1905	Male	Female
EqualAccuracy	Equal Accuracy	0.1007	Female	Male
EqualizedOdds	Equalized Odds	0.0786	Male	Female
EqualOpportunity	Equal Opportunity	0.0786	Male	Female

Example Fit Statistics to Measure Model Performance

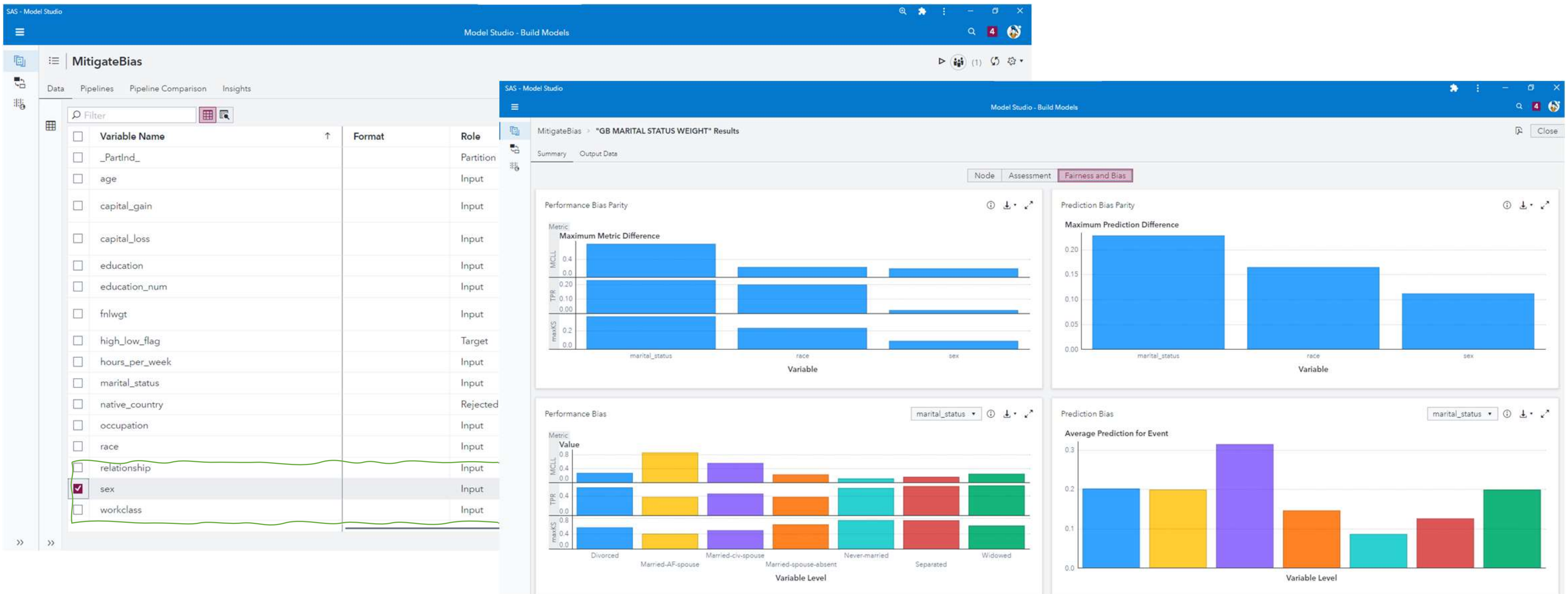
Abbreviation	Fit Statistic Name	Better is:
ASE	Average Squared Error	lower
RASE	Root Average Squared Error	lower
maxKS	Maximum Kolmogorov Smirnov statistic	higher
MCLL	Multi-Class Log Loss	lower
TPR	True Positive Rate	higher
FPR	False Positive Rate	lower
FNR	False Negative Rate	lower
FDR	False Discovery Rate = FP/(FP + TP)	lower
C	Area under ROC	higher
F1	F1 score	higher

Using SAS Model Studio to Assess Bias

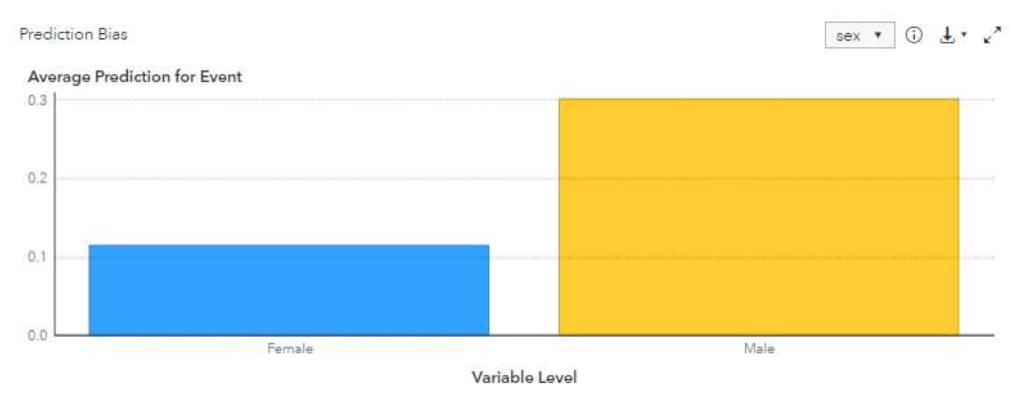
Assessing Bias	
Performance Bias	Fit statistics (e.g., ASE, RASE, MCLL, TPR & KS) grouped by each level of the sensitive variable
Performance Bias Parity	Difference between the variable level with the highest value and the variable value with the lowest value for each of the fit statistics from the Performance Bias Chart
Prediction Bias	Average predicted values for every level (including missing values) of the sensitive variable
Prediction Bias Parity	Difference between the variable level with the highest average prediction for event value and the variable level with the lowest average prediction for event value

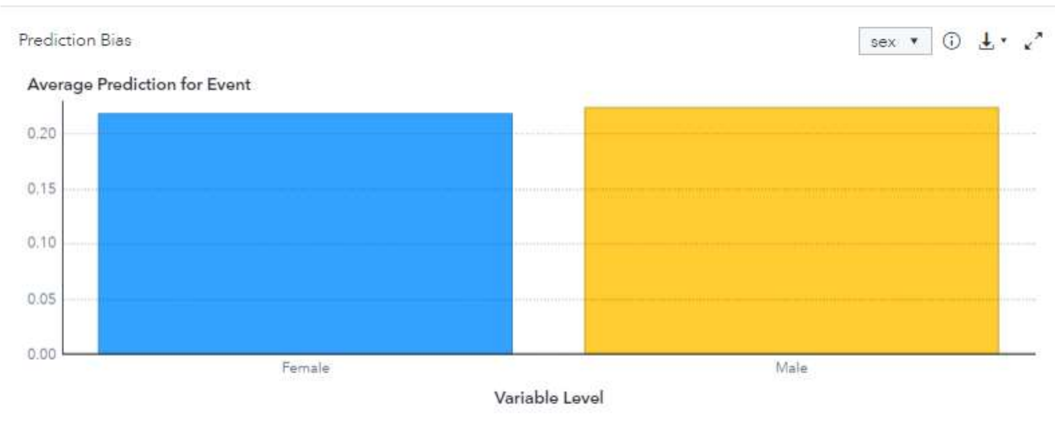
How to Assess Model Fairness with SAS?

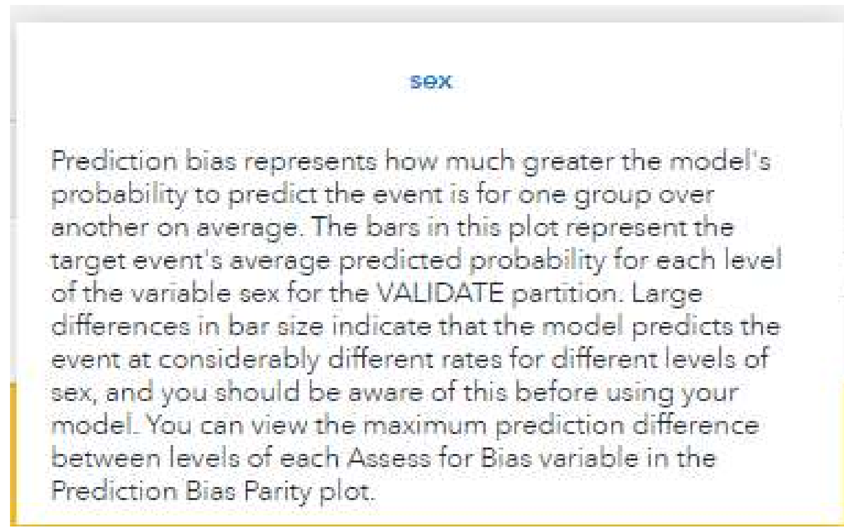
Model Studio



How to mitigate Bias with SAS Viya?







How to Mitigate Model Fairness with SAS?

fairAITools.mitigateBias Action

```
proc cas;
fairAITools.mitigateBias result=rslt /
    biasMetric="PREDICTIVEPARITY",
    bound="625",
    event="1",
    learningRate="0.01",
    maxIters="10",
    predictedVariables={"P_high_low_flag0", "P_high_low_flag1"},
    response="high_low_flag",
    responseLevels={"0", "1"},
    sensitiveVariable="sex",
    table="adult_train",
    tolerance="0.005",
    trainProgram="
        decisionTree.gbtreeTrain result=train_res /
           table=table,
           weight=weight,
           target=""high_low_flag"",
           inputs= {
                    ""age"",
                    ""workclass"",
                    ""fnlwgt"",
                    ""education"",
                    ""occupation"",
                    ""relationship"",
                    ""capital_gain"",
                    ""capital_loss"",
                    ""hours_per_week"",
                    ""native_country""
                    },
           nominals={""workclass"",""high_low_flag"",""education"",
           nBins=50,
```

GROUPMETRICSHISTORY   Table rows: 12 | Columns: 3 of 34 | Rows 1 to 12

Enter expression

	Iteration	Group	PR...
1	1	Female	0.1443
2	1	Male	0.2918
3	2	Female	0.1657
4	2	Male	0.2623
5	3	Female	0.1926
6	3	Male	0.2404
7	4	Female	0.2065
8	4	Male	0.2295
9	5	Female	0.2156
10	5	Male	0.2246
11	6	Female	0.2186
12	6	Male	0.2225

Link to documentation

Demo

Creating Fairness in your AI models with SAS Viya

- Assessing Bias in Visual Analytics
- Assessing and Mitigating Bias in Model Studio

Demo scenario

- Background:
  - https://archive.ics.uci.edu/ml/datasets/census+income
  - Model Objective : The models predicts who ends up in the high (>50.000 usd) or low income <50.000 usd) group(binary variable)
- Data (32.561 obs): Marital_status, education, age,
  - Sensitive variable: sex , race, nationality

Try the demo yourself

- Start up a Viya environment
  - Free 14 day trial
- https://www.sas.com/en_us/trials/software/viya/viya-trial-form.html

Fetch for instance this dataset that I used – or use your own data (but should be non-sensitive, anonymized) :

- https://archive.ics.uci.edu/ml/datasets/census+income
- Follow the steps in Tamara Fischer and Veronique Van Vlasselelaer's webinar on Creating Fair Machine Learning Models

More information

Trustworthy AI
5.10.2023 : webinar on how to put Trustworthy AI into practice
https://www.sas.com/en_us/webinars/trustworthy-ai-using-sas.html
For the insurers
5.10.2023 : webinar on AI and Explainable AI in a ratemaking content
https://www.sas.com/sas/webinars/actuarial-models.html
https://www.sas.com/content/dam/SAS/documents/marketing-whitepapers-ebooks/ebooks/en/a-
comprehensive-approach-to-trustworthy-ai-governance-113518.pdf
About the EU AI Act
The status of AI regulation in the US and elsewhere 24.8.2023
https://open.spotify.com/show/0hSfdnMthaAv3jFIWbxq5e
The Norwegian government position on the EU AI Act
https://www.regjeringen.no/contentassets/939c260c81234eae96b6a1a0fd32b6de/norwegian-position-paper-on-the-ecs-proposal-for-a-regulation-of-ai.pdf
Bias
"Coded Bias“ documentary from netflix https://lnkd.in/eHrzDzXq
https://kjonnsforskning.no/sites/default/files/rapporter/hva_vet_vi_om_kunstig_intelligens_og_likestilling.pdf