BIAS – A LURKING DANGER THAT CAN CONVERT ALGORITHMIC SYSTEMS INTO DISCRIMINATORY ENTITIES

TOWARDS A FRAMEWORK FOR BIAS IDENTIFICATION AND MITIGATION

Thea Gasser
Bern University of Applied Sciences
Bern, Switzerland
thea.gasser@live.com

Eduard Klein (Presenter)
Bern University of Applied Sciences
Bern, Switzerland
eduard.klein@bfh.ch

Lasse Seppänen
Hämeen Ammattikorkeakoulu
Hämeenlinna, Finland
lasse.seppanen@hamk.fi

Presentation for: CENTRIC 2020 - The 13th Int. Conf. on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (October 18, 2020 - October 22, 2020). Porto/Portugal

CONCERNS

2017

FACEBOOK AUTOMATIC TRANSLATION

Choosing wrong translation for a user post leading to the Israeli police interrogating the affected user (Cossins, 2018)

2016

PREDPOL

Targeting a criminal minority unfairly by leading the police to a particular neighborhood
(Cossins, 2018)

2016

COMPAS

Incorrectly judging black defendants more likely to be at higher risk of recidivism while incorrectly judging white defendants more likely as low risk (Larson, Mattu, Kirchner, & Angwin, 2016)

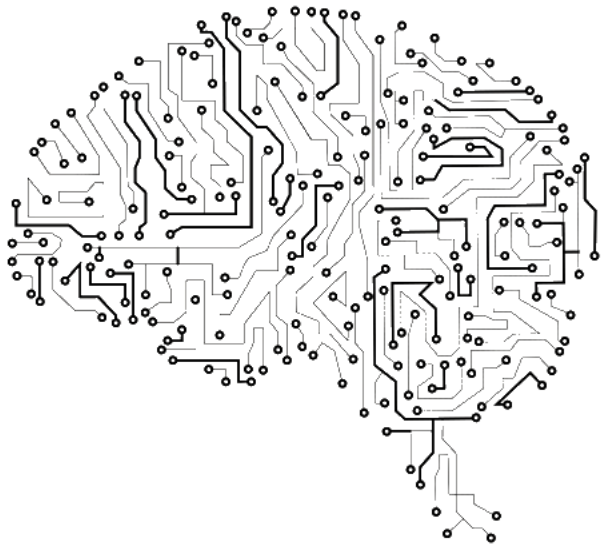
2015

GOOGLE IMAGE RECOGNITION

A software engineer reported that his black friends were classified as “gorillas”
(Vincent, 2018)

TERMINOLOGY

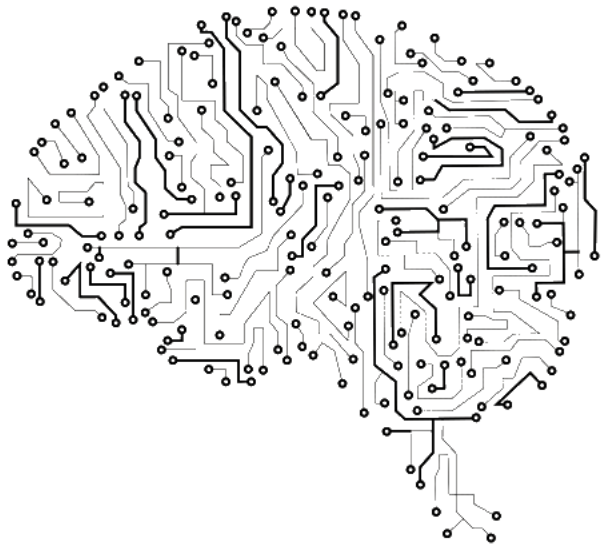
BIAS IN ALGORITHMIC SYSTEMS



UNFAIR

TERMINOLOGY

BIAS IN ALGORITHMIC SYSTEMS

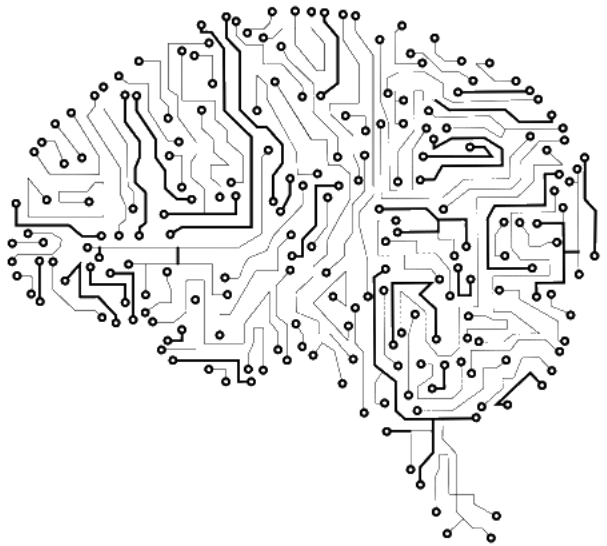


PERSONAL OPINION

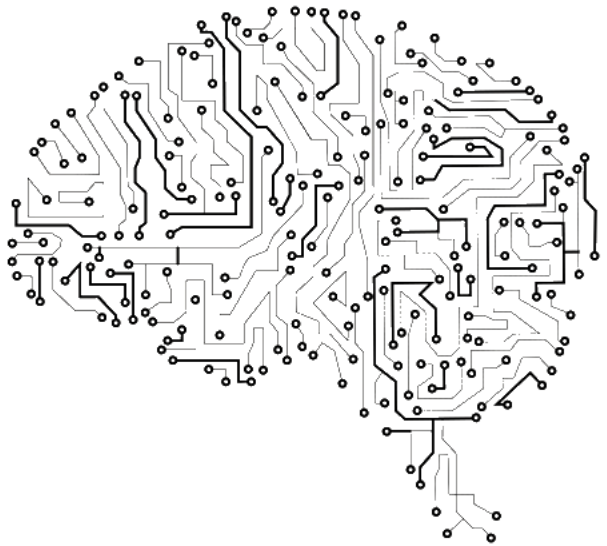
TERMINOLOGY

BIAS IN ALGORITHMIC SYSTEMS

INCORRECTNESS

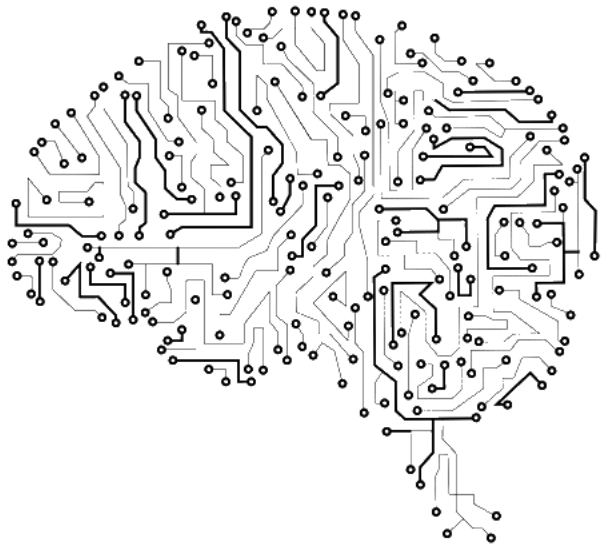


RESEARCH QUESTIONS



- WHAT IS EXPECTED OF AI-SYSTEMS IN RELATION TO HOW HUMANS MAKE DECISIONS?
- HOW IS BIAS THAT AFFECTS HUMAN BEHAVIOUR AND DECISIONS ALSO PRESENT IN ALGORITHMIC SYSTEMS?
- HOW CAN BIAS IN ALGORITHMIC SYSTEMS BE IDENTIFIED?
- WHAT MEASUREMENTS CAN BE TAKEN TO MITIGATE BIAS IN ALGORITHMIC SYSTEMS?

RESEARCH DESIGN

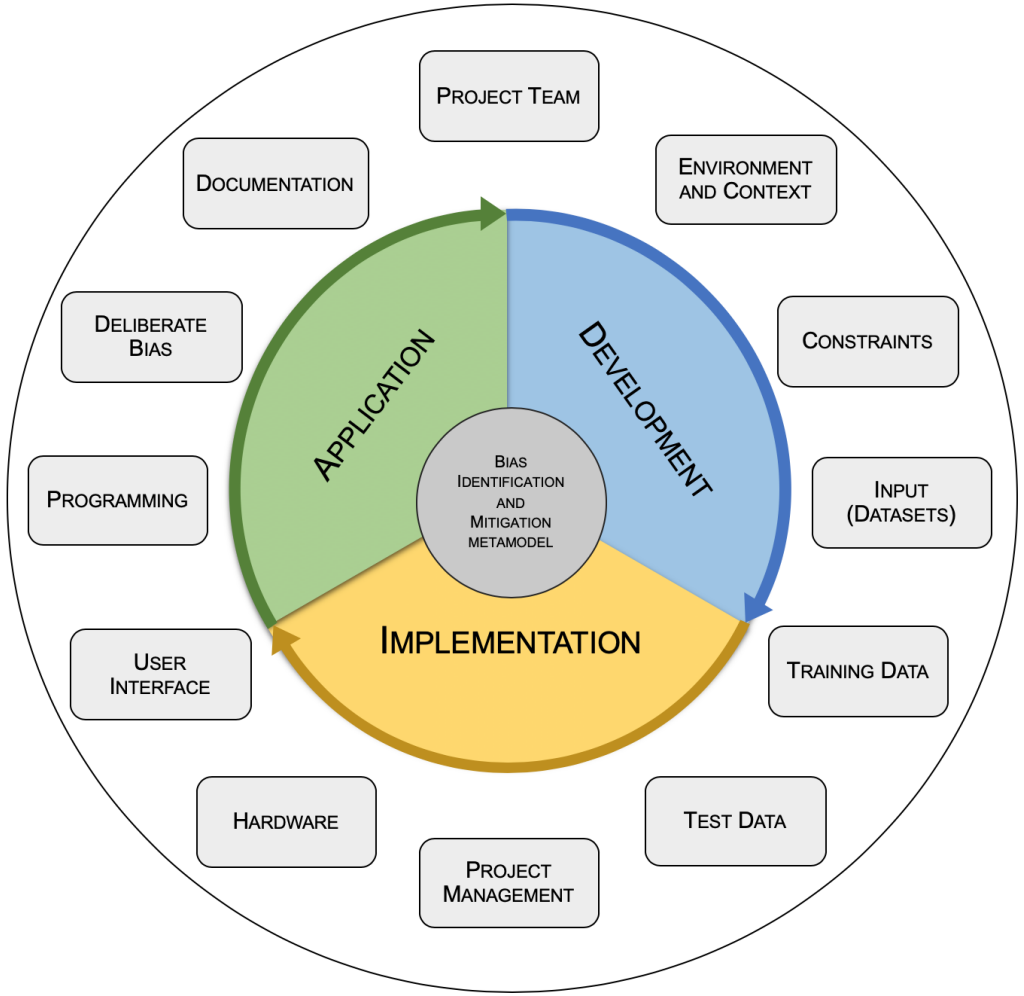


- Extensive and Systematic Literature Research
- Identifying Key Aspects concerning Identification and Mitigation of Bias in (machine learning) Algorithms
- Development of a Framework, useful in project context
- Validation based on Literature
- Validation in real project context

FRAMEWORK

Element	Description/Comments	Yes	No
Project Team			
All project members have had ethical training	Members have a confirmation that they have completed courses or workshops or similar The minimum requirements to consider this element as fulfilled must be defined in the company		
The project team is a cross-functional team, including diversity in ethnicity, gender, culture, education, age and socioeconomic status	The inputs of the same number of men and women, of young and old etc. are included		

CHECKLISTS



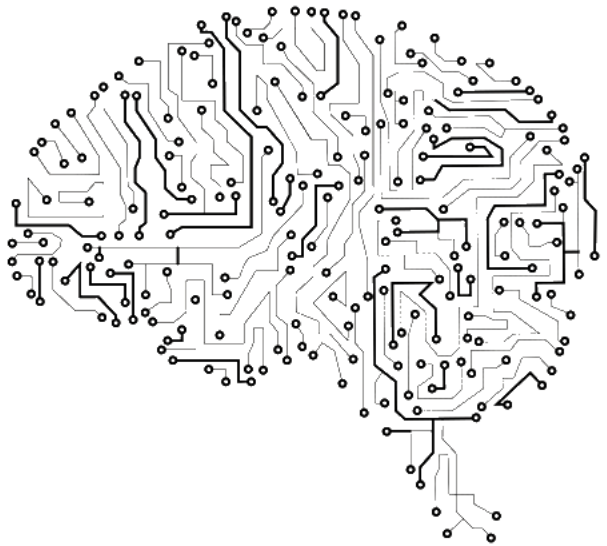
METAMODEL

PROJECT TEAM

Element	Description/Comments	Yes	No
Project Team			
All project members have had ethical training	Members have a confirmation that they have completed courses or workshops or similar The minimum requirements to consider this element as fulfilled must be defined in the company		
The project team is a cross-functional team, including diversity in ethnicity, gender, culture, education, age and socioeconomic status	The inputs of the same number of men and women, of young and old etc. are included		
The project team has representatives from the public as well as the private sector	Exclusions need to be avoided		

TERMINOLOGY

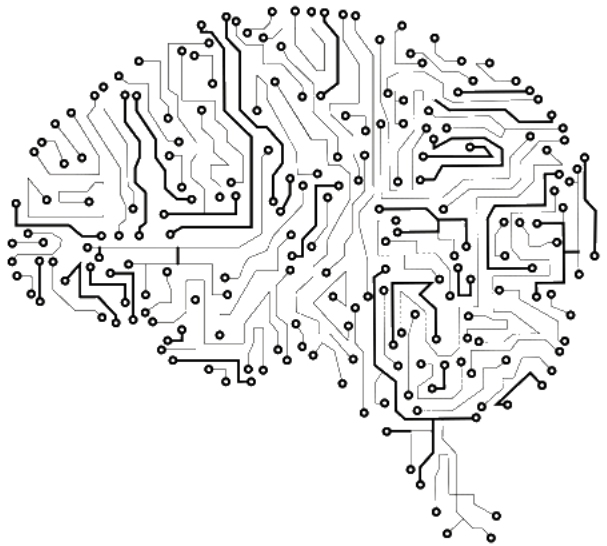
BIAS IN ALGORITHMIC SYSTEMS



UNFAIR

TERMINOLOGY

BIAS IN ALGORITHMIC SYSTEMS

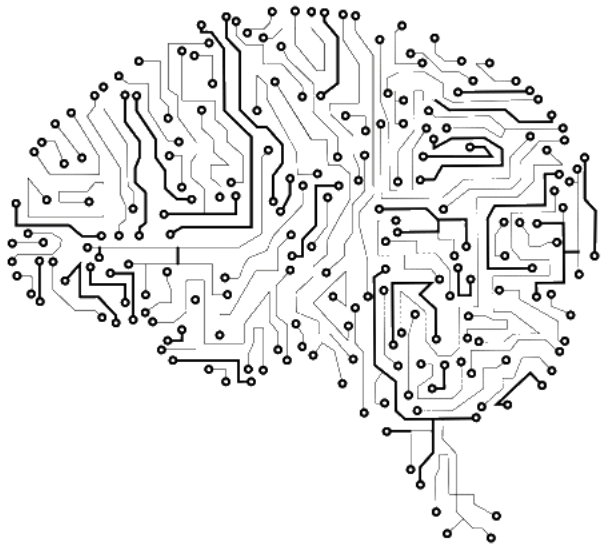


PERSONAL OPINION

TERMINOLOGY

BIAS IN ALGORITHMIC SYSTEMS

INCORRECTNESS



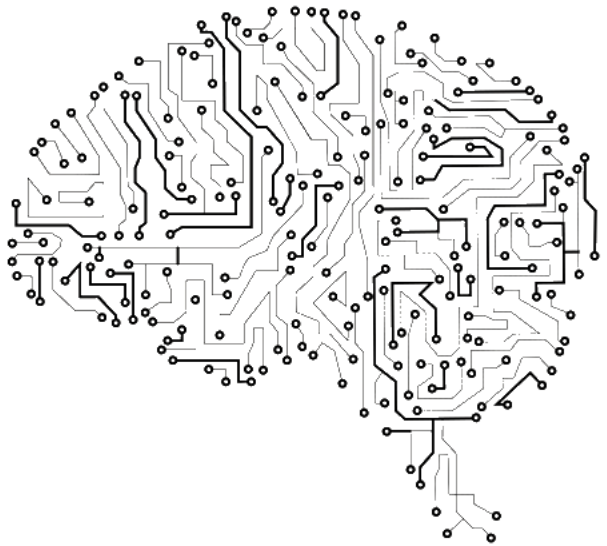
USER INTERFACE

Element	Description/Comments	Yes	No
User Interface			
Visual aspects are determined appropriately	**Text**: the font-style, font-size, font-colour and placement are justified and reflect the intention of the system‘s functionality **Forms /elements**: (e.g. boxes containing text or graphics): colour, size and placements are justified and reflect the intention of the system‘s functionality		
Does visual result representation (alphabetically or random) make any difference (user always choses the results displayed first?)	-		
Is a translation of data/information necessary?	How does the chosen language influence the user‘s perception and interpretation in differet contexts and circumstances?		
Do the information and results become distorted through the aplicatoin of translation?	How is the translation interpreted by the end user?		

PROJECT MANAGEMENT

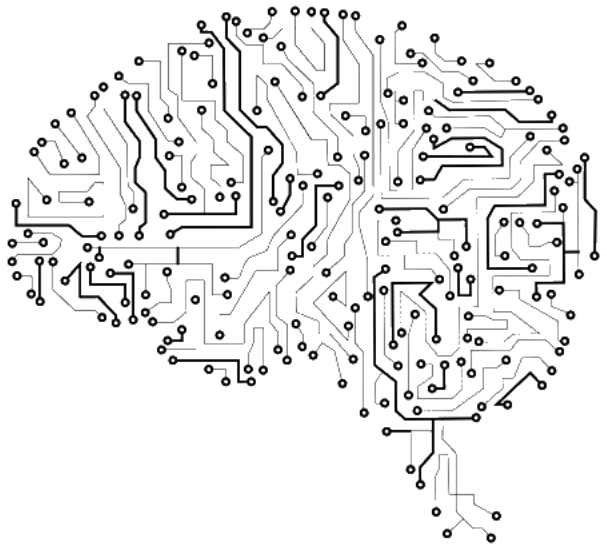
Project Management	
Project management process includes methods that focus on bias issues	- Stakeholder analysis is adjusted for disadvantaged group identification in worst case
Risks concerning bias are assessed and known to each team member	- Risk analysis is adjusted for additional focus on bias and worst-case scenarios provoking to bias
Critical thinking is promoted and demanded at every stage of the system creation process	- How would changes to a data point affect the model's prediction? - Does it perform differently for various groups? For example, historically marginalised people? - How diverse is the dataset I am testing my model on? - Is the system context the one the system was intended to? - Can the outcome/result/system recommendation be justified? - How diverse is the dataset I am testing my model on? - Does it perform differently for various groups–for example, historically marginalized people? - How would changes to a data point affect my model's prediction?
Perspectives are changed continuously	- Different points of views ensure identification of hidden as-

FAIRNESS



- FAIRNESS HARD TO DEFINE
- FAIRNESS DEPENDS ON THE CONTEXT AND VIEW
- DELIBERATE BIAS

OUTLOOK

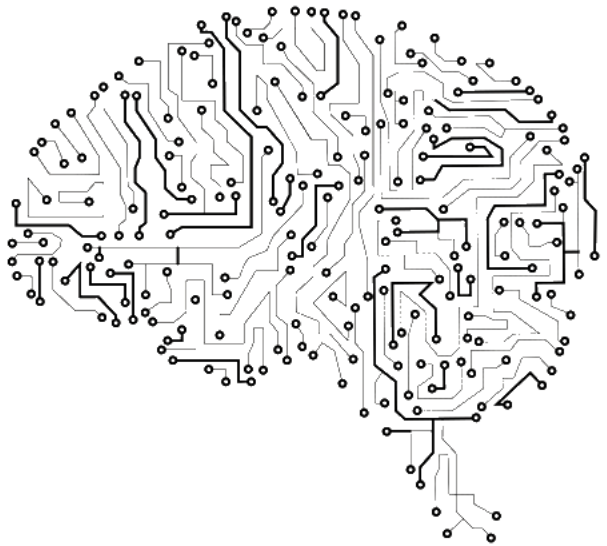


- CONTRIBUTING TO AI-SAFETY
- APPLICATION OF FRAMEWORK IN REAL PROJECT CONTEXT
- ADJUSTMENTS OF FRAMEWORK DEPENDING ON CONTEXT

THANK YOU FOR LISTENING

QUESTIONS ?

BIBLIOGRAPHY



Cossins, D. (2018). Discriminating algorithms: 5 times AI showed prejudice. https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/

Feuerriegel S., Dolata M., Schwabe G. (2020). "Fair AI". Bus. Inf. Syst. Eng., pp. 379-384. https://link.springer.com/article/10.1007/s12599-020-00650-3

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Vincent, J. (2018, January 12). Google ‘fixed’ its racist algorithm by removing gorillas from its image-labeling tech. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai