UNIVERSITÉ **CÔTE D'AZUR**

License for this content: CC BY-NC-SA

Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material

| Share your work | Use & remix | What We |

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

NonCommercial — You may not use the material for commercial purposes.

ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

As long as we have good data, are we going to get there?...

Spoiler: Not necessarily

LET'S TAKE A STEP BACK: WHAT TASKS ARE WE AIMING FOR?

ML Deployment Failures in High-Stakes Scenarios

- AI systems have already been deployed in high-stakes scenarios and have failed with terrible consequences:
  - AI-based unemployment benefit fraud detectors have left (innocent) people without income, paralyzed people have had their home help cut in half, …
- And these failures disproportionately discriminate against disadvantaged socio-demographic groups. The African-American community has been overly targeted by system failures:
  - used to identify criminals and predict recidivism rates
  - Low-income people were wrongly identified as less in need of medical assistance
  - more likely to abuse children
  - Women were identified as less attractive to recruit

[1] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst, "The Fallacy of AI Functionality," in 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea, June 2022, pp. 959–972, ACM.
[2] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt, "Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning," in Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

Rightly or wrongly?

- Let's be clear: these decisions were wrong

- Child Maltreatment Prevention:
  - Strong oversampling of working-class and families of color, subjecting poor parents and children to more frequent surveys

- Hospital Bed Allocation:
  - Care needs quantified by individual healthcare spending!

- Recruitment of women:
  - Ground truth (the target of the system) is made up of a history of biased human decisions
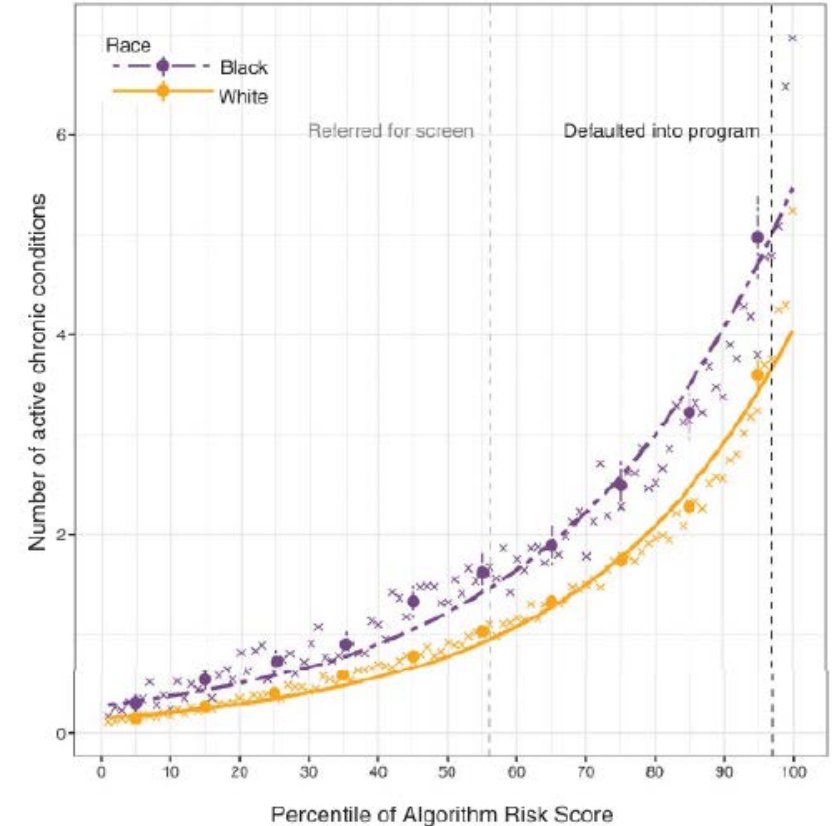
Taken from Obermeyer et al. [7]. The figure shows that at a given risk score produced by the algorithm, Black patients are considerably sicker than White patients.

[1] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst, "The Fallacy of AI Functionality," in 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea, June 2022, pp. 959–972, ACM.

Facial Recognition

- Buolamwini and Gebru 2018: inequity in recognition following an intersection of factors: gender + skin color
- Pressure persists despite systematic deployment failures
  - French municipalities under pressure
  - 2022, EDPB: call some interdictions for facial recognition for the police
  - 2024, Olympic Games: lobbying by France in the AI Act
- And that's not all:
  - Light skin face over-representation in face datasets for facial recognition
  - Over-representation of Western world objects for the recollection of objects
  - Male Pronouns and Masculine Nouns for Named Entity Recognition
- → Realizing that datasets reflect the dominance of intersectional groups, and that this impacts models.

2015: Misclassification of minorities

2020: Biased super-resolution
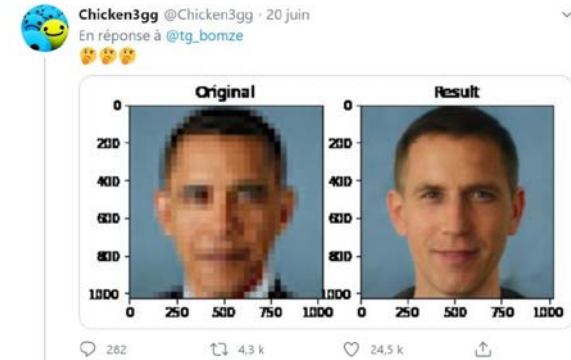
2018: Intersectional bias in face tech.
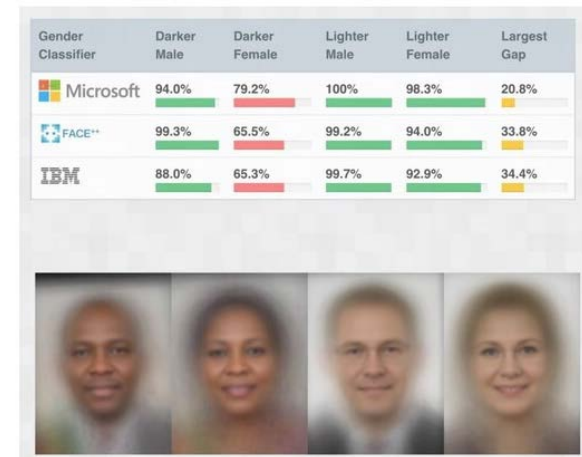
Source: *Buolamwini & Gebru (2018)*

[1] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst, "The Fallacy of AI Functionality," in 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea, June 2022, pp. 959–972, ACM.[1]

[2] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *Patterns*, Nov. 2021.

[3] J. Hourdeaux. JO 2024 : l'expérimentation de la vidéosurveillance algorithmique inquiète. Mediapart, janvier 2023.

univ-cotedazur.fr

- What assumptions should be made to reduce the complexity of the world in order to model the problem? What data should be used?

- Should we try to correct them? Can we?

- Should we question the automation itself, whatever the method, of these sensitive decisions?

Humans involved

- Every dataset therefore involves humans:
  - those who decide the target task,
  - those who decide how to collect data samples,
  - those who decide the annotation guidelines,
  - those who decide who annotates,
  - those who are assigned the annotation work,
  - those whose personal data is used.
- This simple fact leads to limitations and biases in any data-driven approach, no matter how massive.

A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *Patterns*, Nov. 2021

Humans involved

- To understand the connection between our ML research practices and the social and structural problems pervading datasets, let us introduce the guidelines of Paullada et al. for ML practitioners when we:
  - (1) define a problem to be tackled with ML,
  - (2) create or choose existing data to use
  - (3) analyze the model performance and envision real-world deployment.

A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *Patterns*, Nov. 2021

Define a problem to be tackled with ML

- Tasks can be defined abstractly ("intentionally") as a problem statement (e.g., object recognition, speech-to-text translation) or "extensionally", that is instantiated by a learning problem made of a dataset of (input, output) pairs and an evaluation metric (e.g., top-1 accuracy)

- One must first analyze the intentional definition of the task and the mapping we can foresee between input and output.

Which tasks to tackle with ML?

- What is the correspondence between the input and the output?
- Face → Sexual Orientation or Employability
  - Pseudo-scientific task based on assertions of essentialism of human traits
- Students' Short Text Responses → IQ score
  - The responsibility lies in (i) legitimizing the IQ score as a reasonable quantity, (ii) predicting IQ with an ML approach, and (iii) assuming that IQ can be predicted from short text responses.
- Prediction of recidivism with the COMPAS system in the US justice system:
  - White violent recidivists 63% more likely to be misclassified as low risk than Black recidivists
  - How? Race is not an input variable, but 137 questions like "Was one of your parents ever sent to jail or prison?" "How many of your friends/acquaintances are taking drugs illegally?" and "How often did you get in fights while at school?"
- A scientist must dare to ask: should recidivism be predicted? Should recidivism be predicted in order to inform legal decisions about individuals?
- If we want to give everyone equal opportunities, whatever their social background, are there any acceptable characteristics on which to base the prediction of recidivism?
- Social determinism underpins tasks tackled with ML (example: predicting student success? To inform Parcoursup decisions?)

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.," Propublica, 2016.
[2] Timnit Gebru and Emily Denton, "Tutorial on Fairness Accountability Transparency and Ethics in Computer Vision at CVPR 2020," https://sites.google.com/view/fatecv-tutorial/home, 2020.
[3] Cathy O'Neil. Weapons of math destruction. 2016.

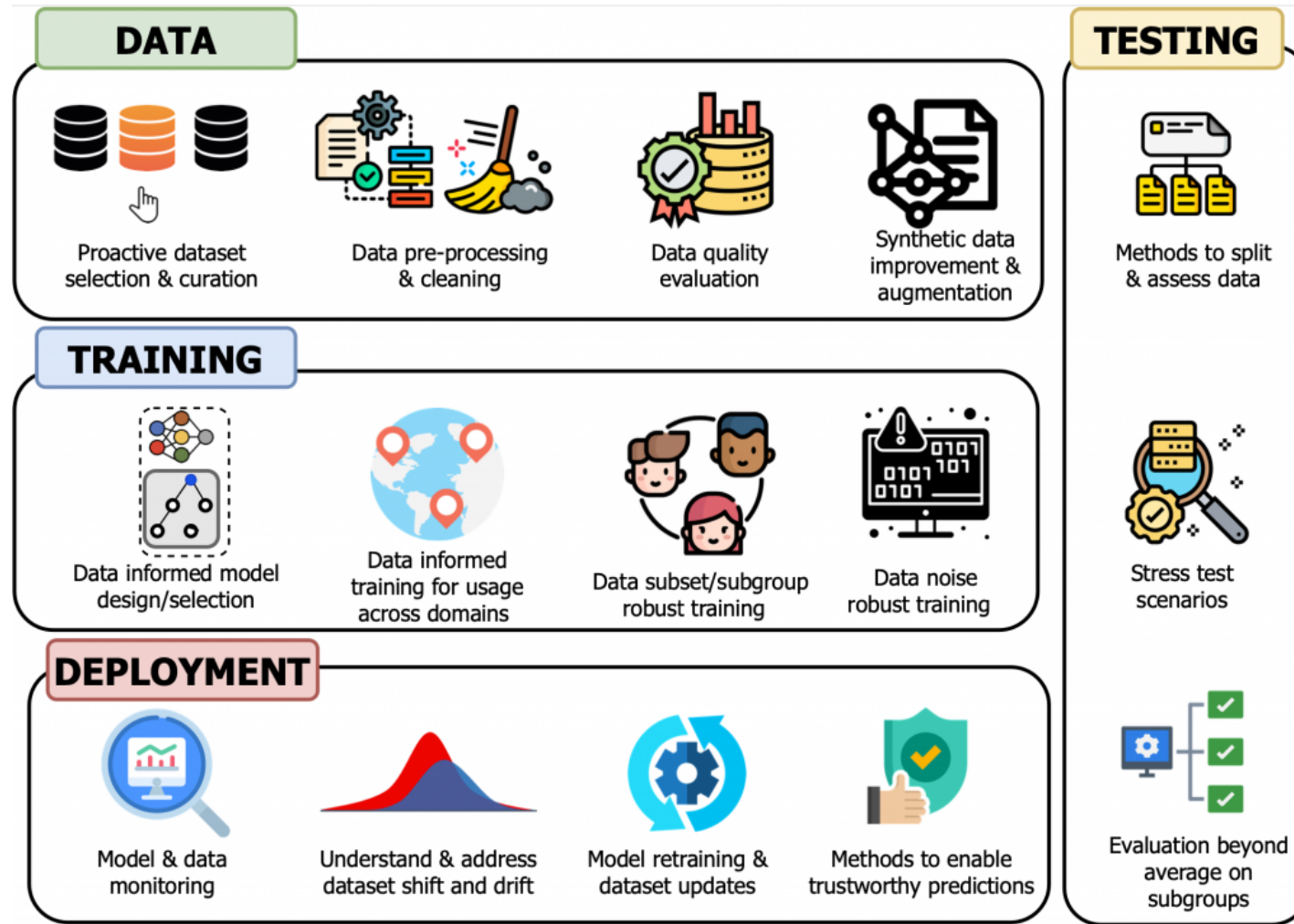# Essentialism, bias, stereotype threat... and ML?

- Pre-recruitment evaluation:
  - Answers, games, short video of the candidate → employability
  - Beyond questionable input-output mapping, another crucial problem in discrimination:

"Cognitive assessments have imposed adverse impacts on minority populations since their introduction into mainstream use. Critics have long contended that observed group differences in test outcomes indicated flaws in the tests themselves, and a growing consensus has formed around the idea that while assessments do have some predictive validity, **they often disadvantage minorities despite the fact that minority candidates have similar real-world job performance to their white counterparts**. **The American Psychological Association (APA) recognizes these concerns as examples of "predictive bias" (when an assessment systematically over- or under-predicts scores for a particular group)** [...] Disparities in assessment outcomes for minority populations are not limited to pre-employment assessments. In the education literature, the adverse impact of assessments on minorities is well-documented. This has led to a decades-long line of literature seeking to measure and mitigate the observed disparities "
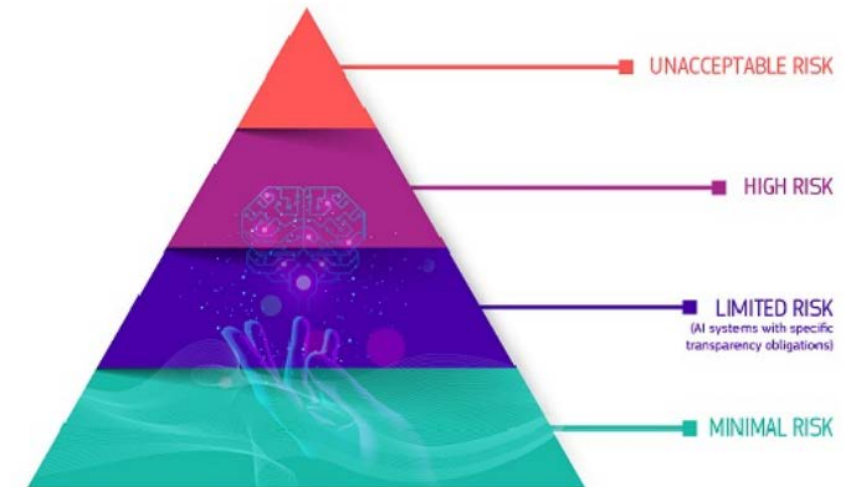
Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," Conference on Fairness, Accountability, and Transparency (ACM FACCT), 2020.

From Model-centric AI to Data-centric AI

From Prof. Michela Van Shaar website: https://www.vanderschaar-lab.com/data-centric-ai/

Our responsibility in the chain

- Catherine Tessier, ONERA's Scientific Integrity and Research Ethics Referent, member of the National Pilot Committee on Digital Ethics
  - "There cannot be an ethical algorithm, but there needs to be an ethic of autonomy"
  - The moral machine is a delusion that hides from us the real choices that scientists and society must be able to consider honestly, outside of the algorithm.
- Jacobsen: emphasizes that "When assessing whether a task is solvable, we first need to ask: should it be solved? And if so, should it be solved by AI?"
- For these reasons: EU AI act
  - AI systems used in the administration of justice, to control access to education and employment are now classified as high-risk systems in the latest EU AI Regulation.

UNACCEPTABLE RISK

HIGH RISK

LIMITED RISK
(AI systems with specific transparency obligations)

MINIMAL RISK

The risk-based approach defines four levels of risk. High-risk AI systems include those "that can determine a person's access to education and career path", "used in employment, worker management and access to self-employment", "used in the administration of justice and democratic processes".

[1] Jörn-Henrik Jacobsen, Robert Geirhos, and Claudio Michaelis, "Shortcuts: Neural networks love to cheat," The Gradient, 2020.
[2] Catherine Tessier. Il n'y a pas de « décision autonome éthique » mais nécessité d'une éthique de l'« autonomie ». La revue de la société savante de l'Aéronautique et de l'Espace, Fév. 2021.

whistleblower

- https://www.linkedin.com/posts/luca-bertuzzi-186729130_aiactfinalfour-column21012024pdf-activity-7155091883872964608-L4Dn?utm_source=share&utm_medium=member_android
- https://www.linkedin.com/posts/dr-laura-caroli-0a96a8a_ai-act-consolidated-version-activity-7155181240751374336-B3Ym/?utm_source=share&utm_medium=member_desktop

2021/0106 (COD)

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

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular Articles 16 and 114 thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the European Economic and Social Committee[1],

Having regard to the opinion of the European Central Bank[2],

Having regard to the joint opinion of the European Data Protection Board and the European Data Protection Supervisor,

Having regard to the opinion of the Committee of the Regions[3],

Acting in accordance with the ordinary legislative procedure,

Whereas:

(1)   The purpose of this Regulation is to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, placing on the market, putting into service and the use of artificial intelligence systems in the Union in

[1]   OJ C […], […], p. […].
[2]   Reference to ECB opinion
[3]   OJ C […], […], p. […].

UNIVERSITÉ CÔTE D'AZUR

Any Question?