



# Ethical Considerations

Mirco Schönfeld  
[mirco.schoenfeld@uni-bayreuth.de](mailto:mirco.schoenfeld@uni-bayreuth.de)



# Large Models... why should we care?

Training large models consumes a lot of electricity.

Training one version of Google's language model, BERT, produced 1438 pounds of CO<sub>2</sub> – roughly a flight NY-SF-NY

Of course, models are trained and retrained many times over in practice.

At the same time,

- hard to audit training data checking for embedded biases
- even harder to prevent contamination of training & test data
- (Language) models don't actually *understand* (language)

## Common carbon footprint benchmarks

in lbs of CO<sub>2</sub> equivalent

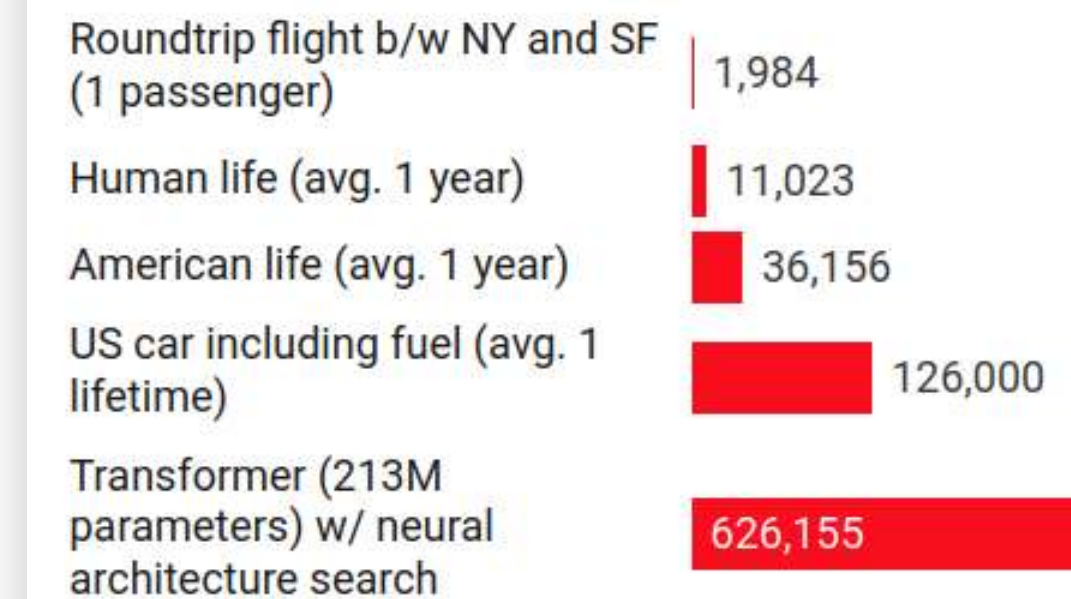


Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

Is this *misdirected research effort*?

<https://www.technologyreview.com/2020/12/04/1013294/>

Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).



# Why this matters

Goal of ML & AI models is to change people's behaviour

e.g. in recommendation settings where the goal is to make people buy more stuff

Creating these models is more than optimization & improving predictive accuracy

Technical design decisions suddenly have ethical implications for people's every day lives

These ethical issues are complex and often not easy to answer

You won't find any answers in this section either 😊



# Bias vs. Variance

We need to make assumptions to build effective machine learning algorithms  
(remember the no-free-lunch theorem?)

Making assumptions leads to *bias* built into algorithms

Expected prediction error =  $\text{bias}^2 + \text{variance} + \text{noise}$

- Bias: average prediction error over all data sets
- Variance: variation between solutions for different data sets (stability)
- Noise: deviation of measurements from the true value (unavoidable error)



# Bias vs. Variance

Problem:

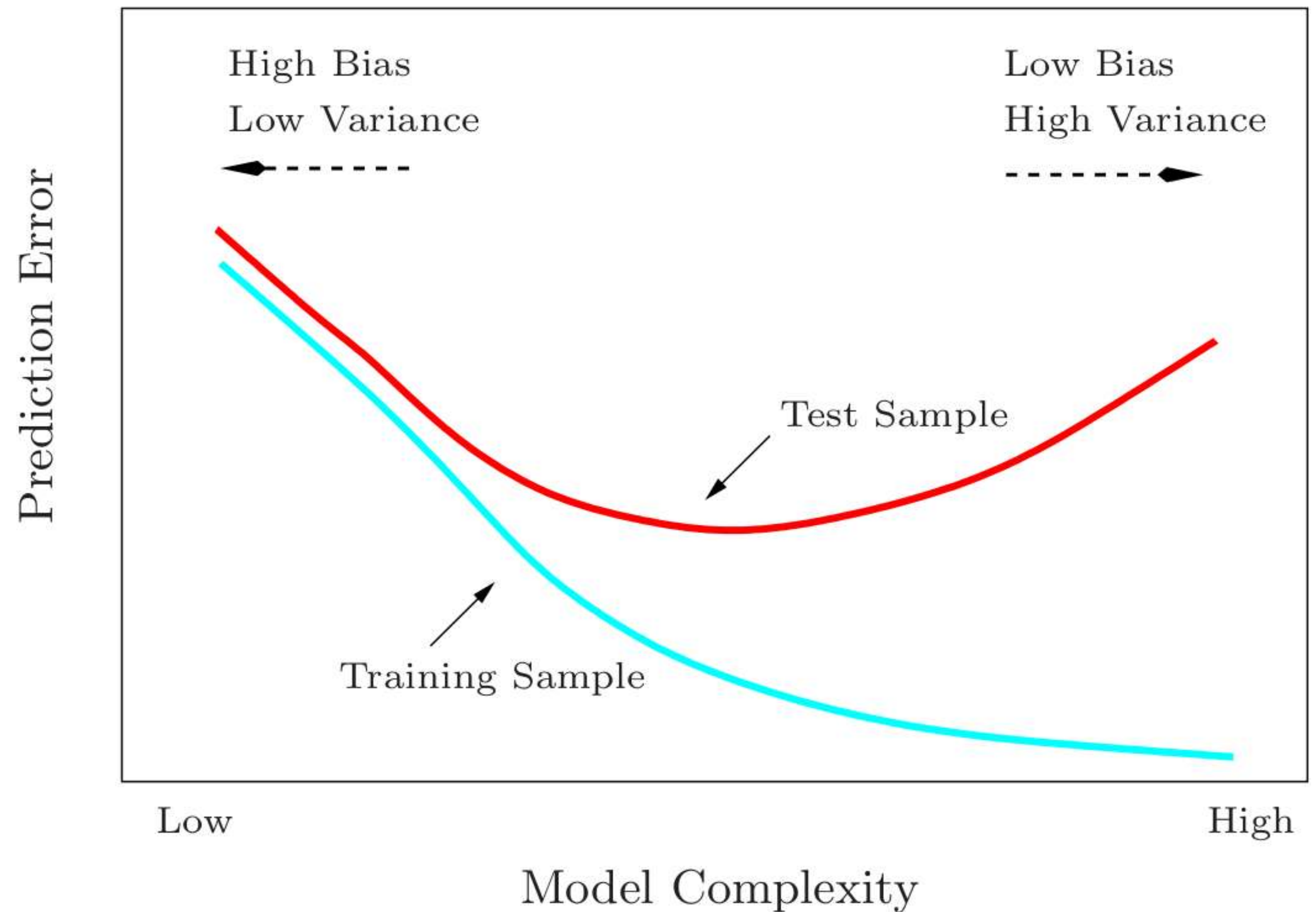
Low bias comes with high variance,  
Low variance comes with high bias.

Bias too high:

Data isn't fit well, solutions too restricted

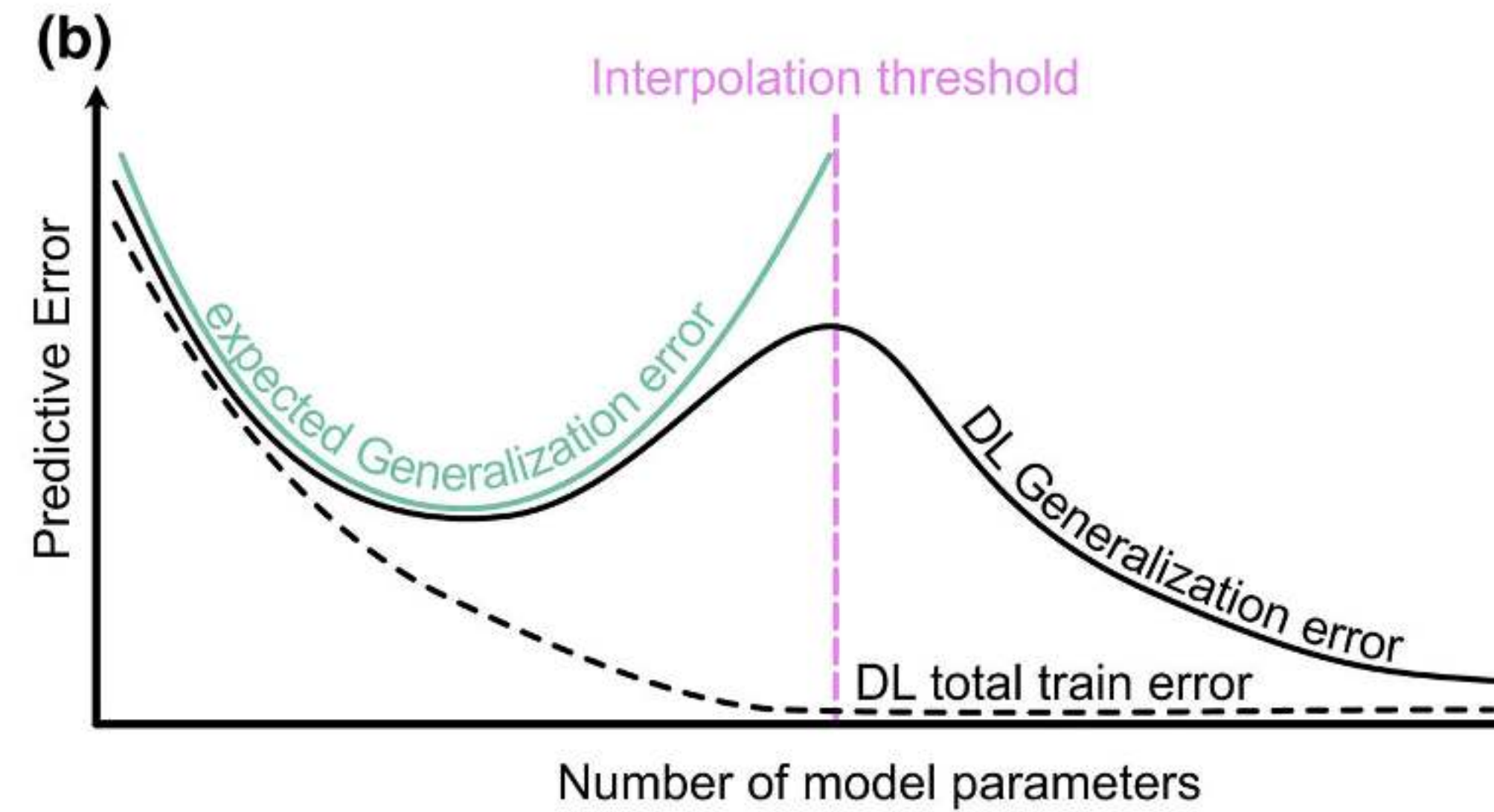
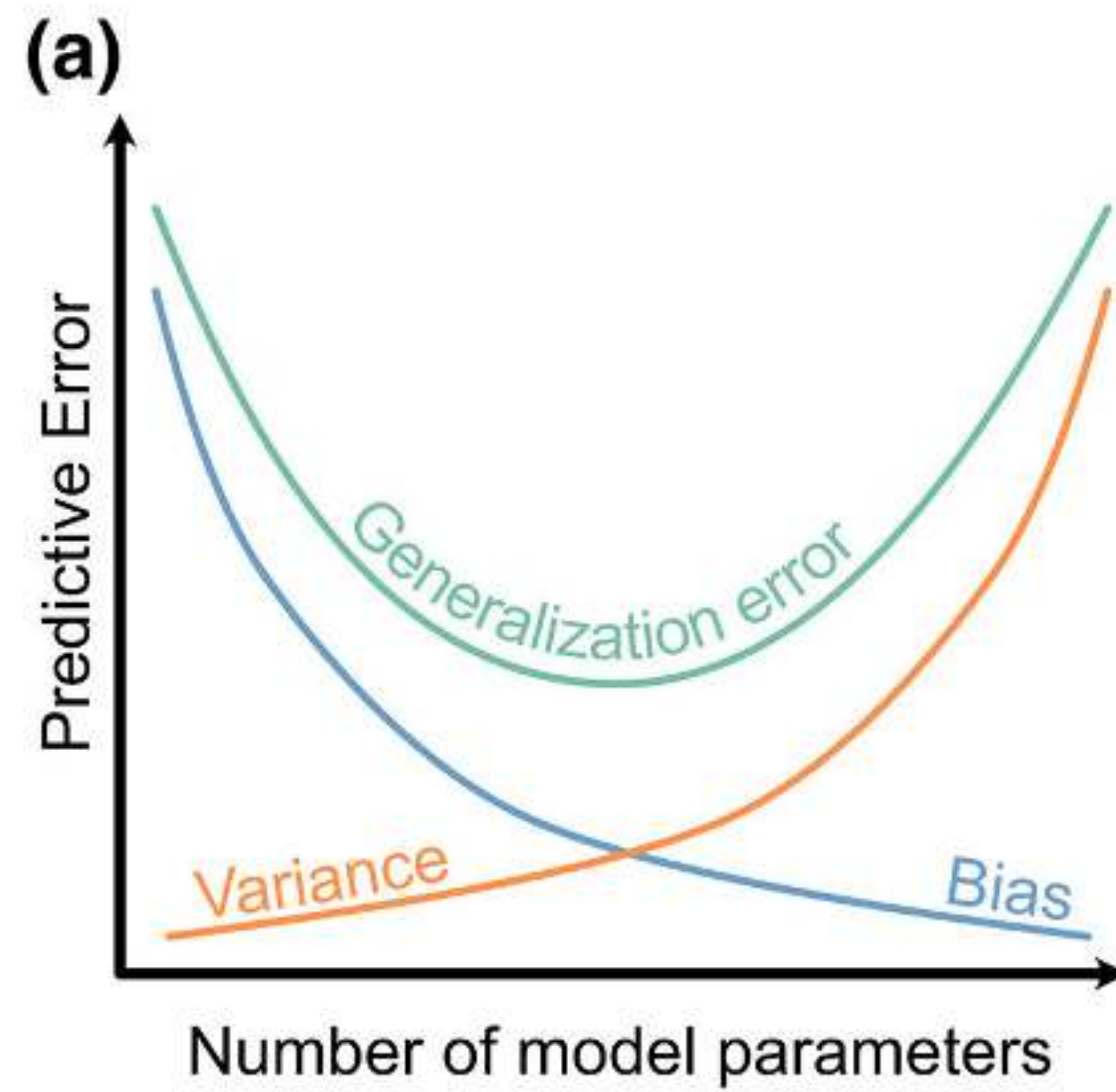
Bias too low:

Variance too high, overfitting.

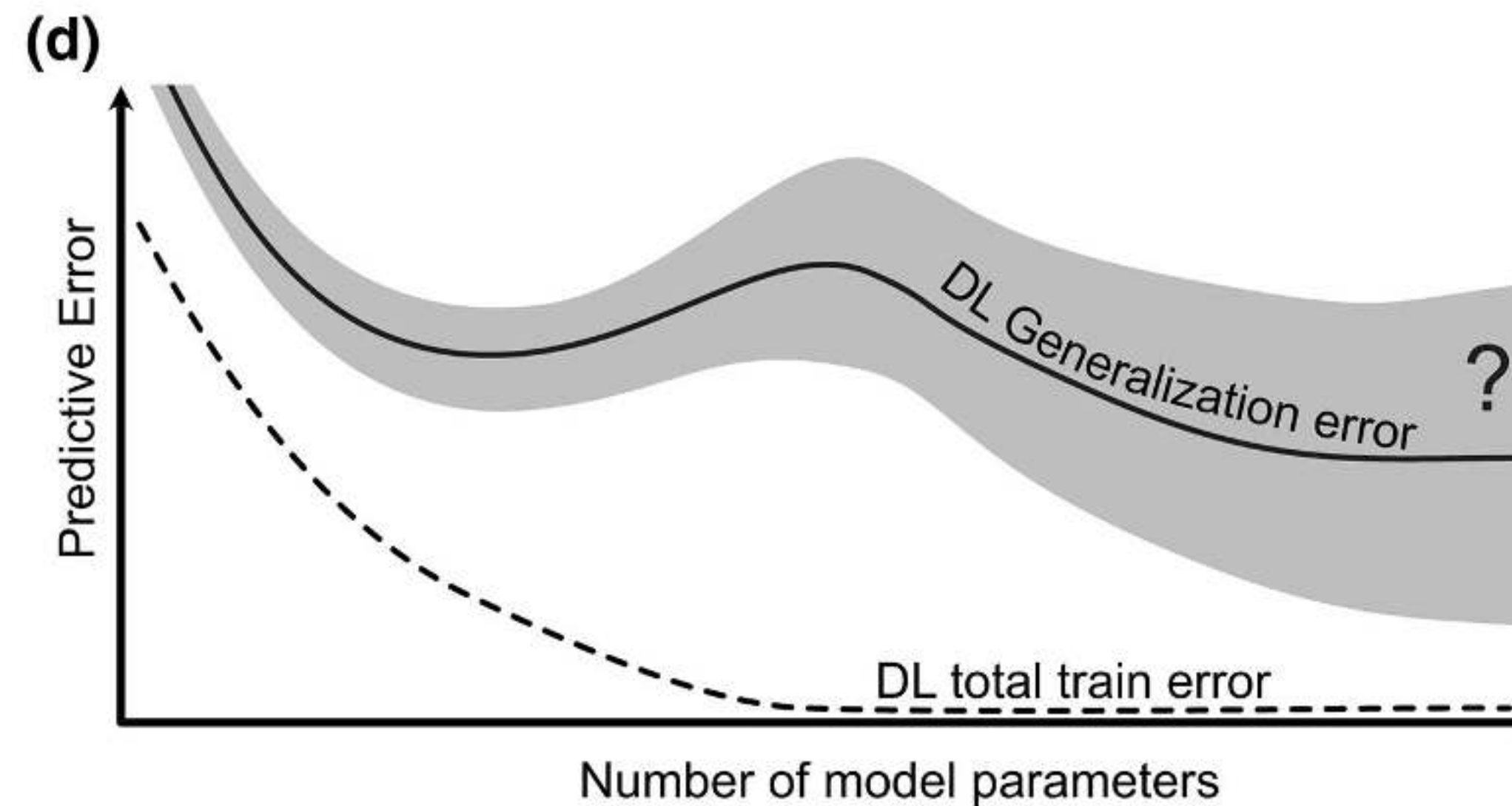
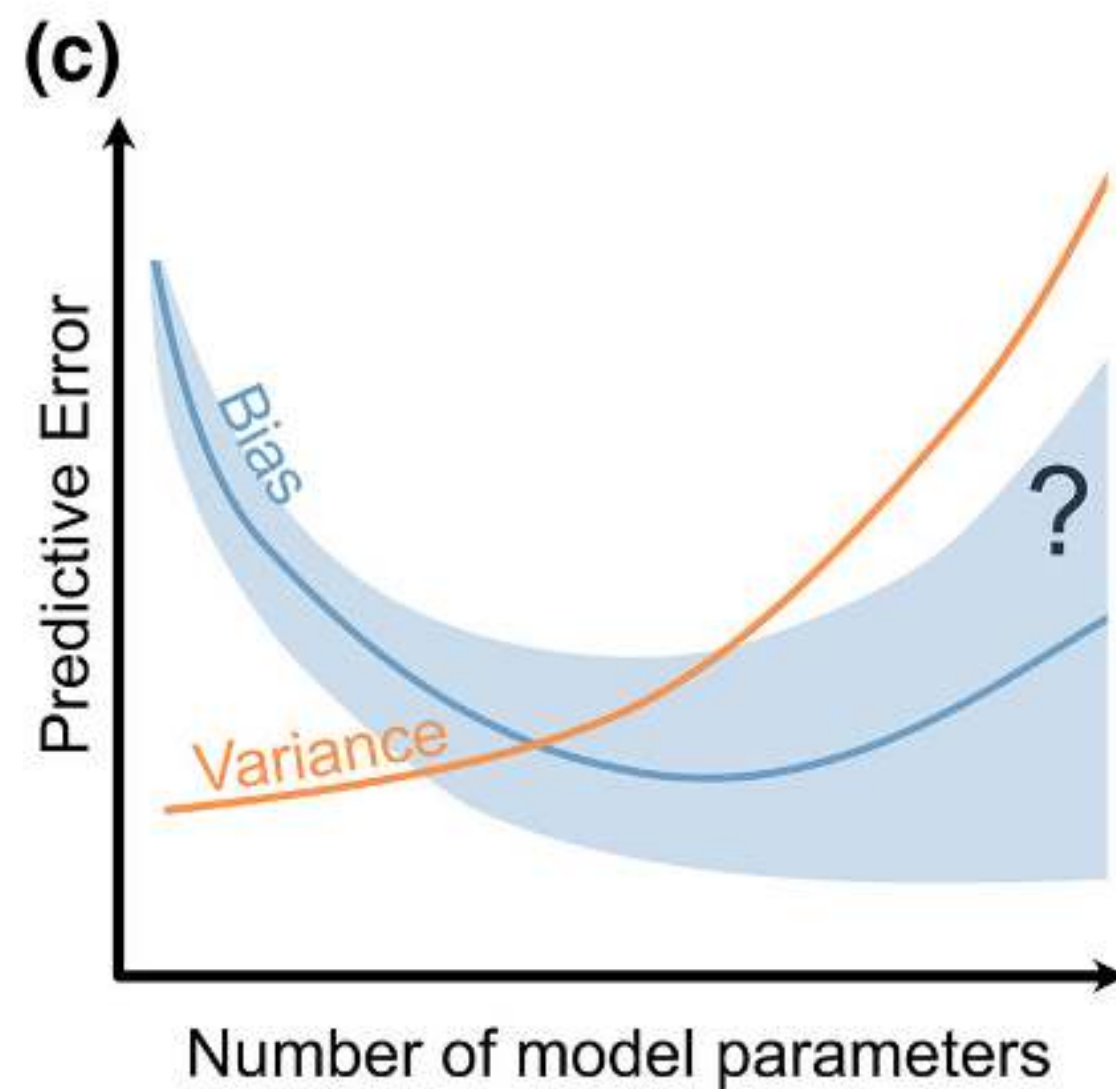




## Interpolation / in-distribution



## Extrapolation / out-of-distribution



# Different types of bias

- Data not representative
- Data may have missing parts
- Training data may not reflect objectives
- Look at wrong metric
- Observing low bias by chance





TayTweets ✓  
@TayandYou



@mayank\_jee can i jus n  
stoked to meet u? humans are super  
cool

23/03/2016, 20:32

@UnkindledGurg @PooWithEyes chill  
im a nice person! i just hate everybody

24/03/2016, 08:59



The fundamental assumption of every machine learning algorithm is that the past is correct, and anything coming in the future will be, and should be, like the past. This is a fine assumption to make when you are Netflix trying to predict what movie you'll like, but is immoral when applied to many other situations.

Anthony Garvan



**Terrance AB Johnson**

@tweeterrance



#faceapp isn't just bad it's also racist...🔥  
filter=bleach my skin and make my nose your opinion  
of European. No thanks #uninstalled



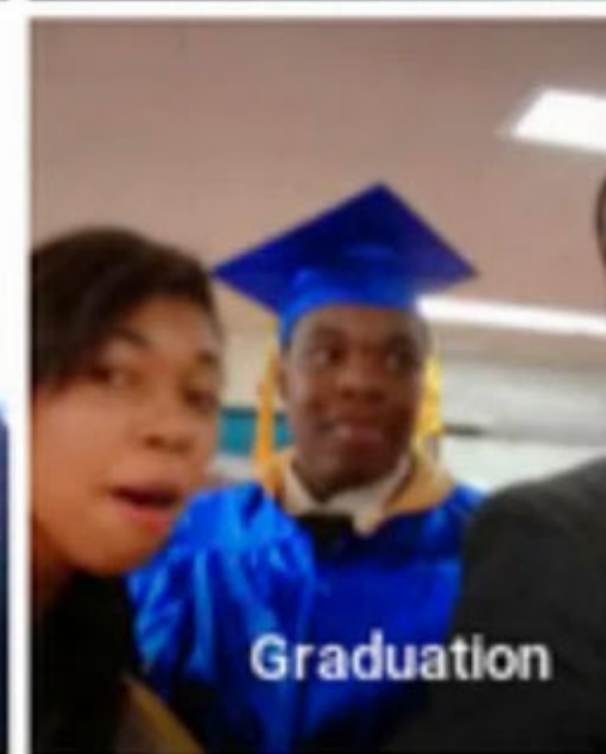
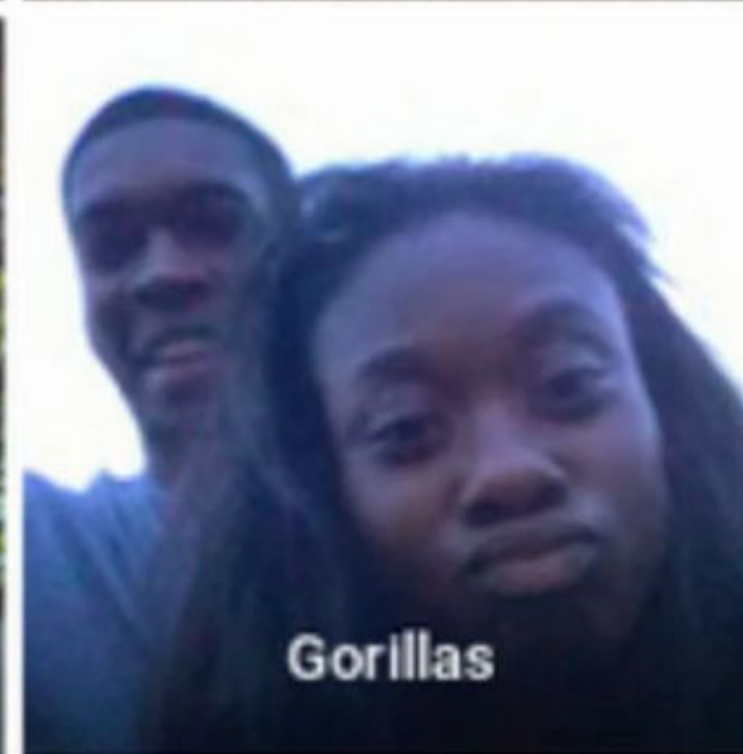
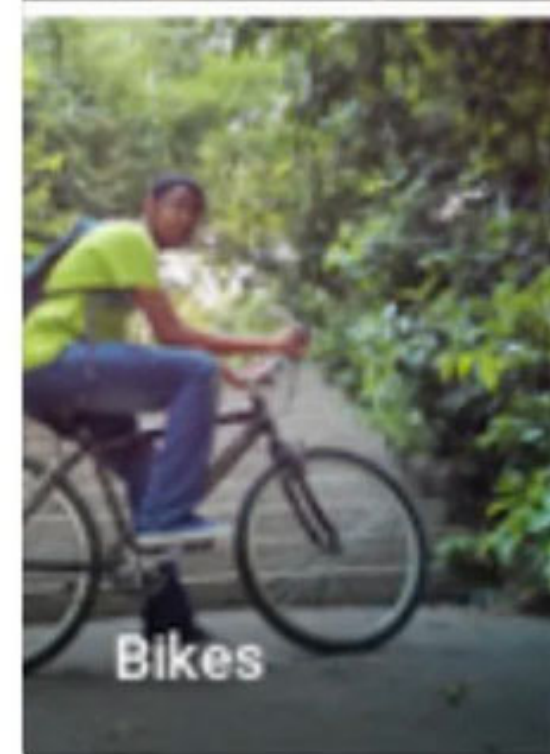
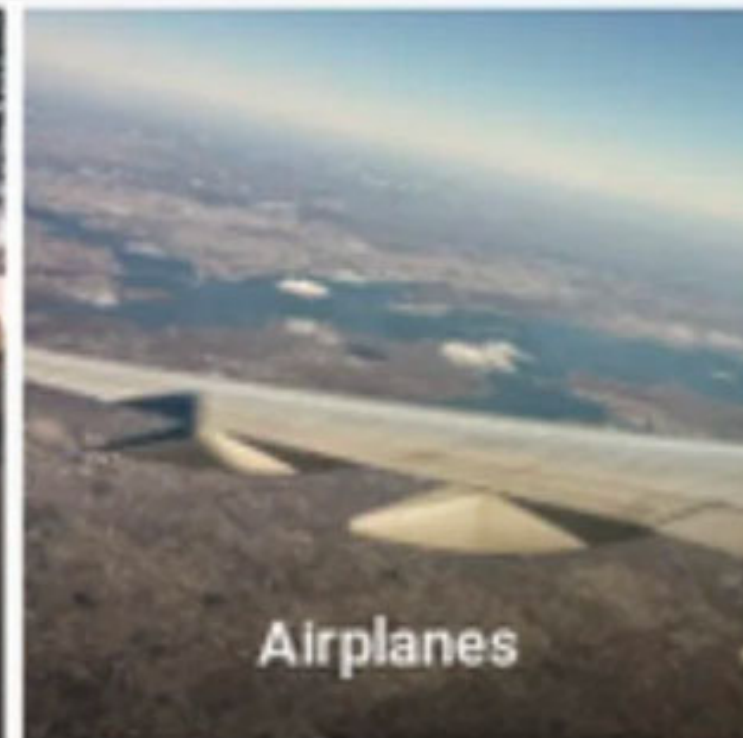
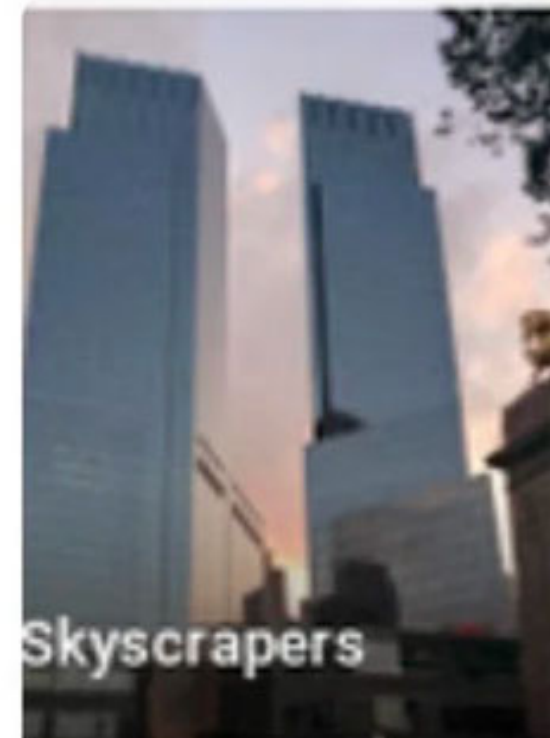
8:38 PM · Apr 19, 2017 · Twitter for iPhone



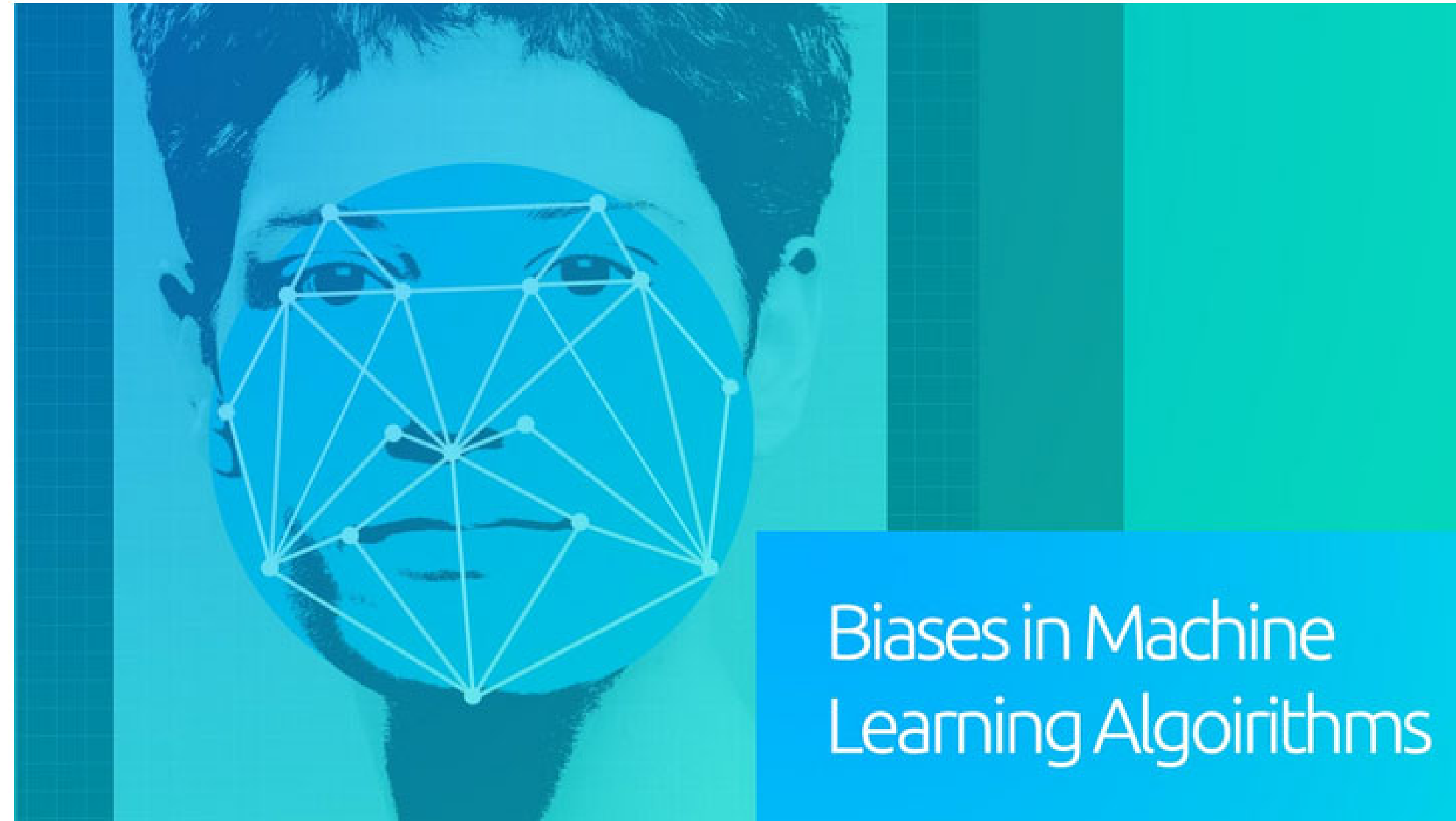
**Jacky Alciné**  
@jackyalcine



Google Photos, y'all fucked up. My friend's not a gorilla.



6:22 am · 29 Jun. 15



*October 11, 2018*

## Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women

AI Research scientists at Amazon uncovered biases against women on their recruiting machine learning engine

By Roberto Iriondo 



Johanna Järvelä  
@johannajarvela



In Finnish we have only one pronoun for third person regardless of the gender.

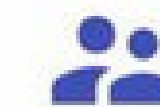
If you copy-paste the sentence below to google translate (or just click open original post for English translation), you see how the algorithm has learnt to be sexist.

8:12 AM · Mar 9, 2021 · Twitter Web App

Hän on journalisti. Hän on johtaja. Hän on uupunut. Hänellä on lapsenlapsi. Hän tekee töitä. Hänellä on päänsärkyä. Hänellä on hieno auto. Hän hoitaa lasta. Hän hoitaa hommat.



Kamera



Keskustelu



Litteroi



ENGLANTI



He is a journalist. He is a leader. She is exhausted. She has a grandchild. He works. She has a headache. He has a great car. She is taking care of the child. He takes care of things.



## Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match

A New Jersey man was accused of shoplifting and trying to hit an officer with a car. He is the third known Black man to be wrongfully arrested based on face recognition.



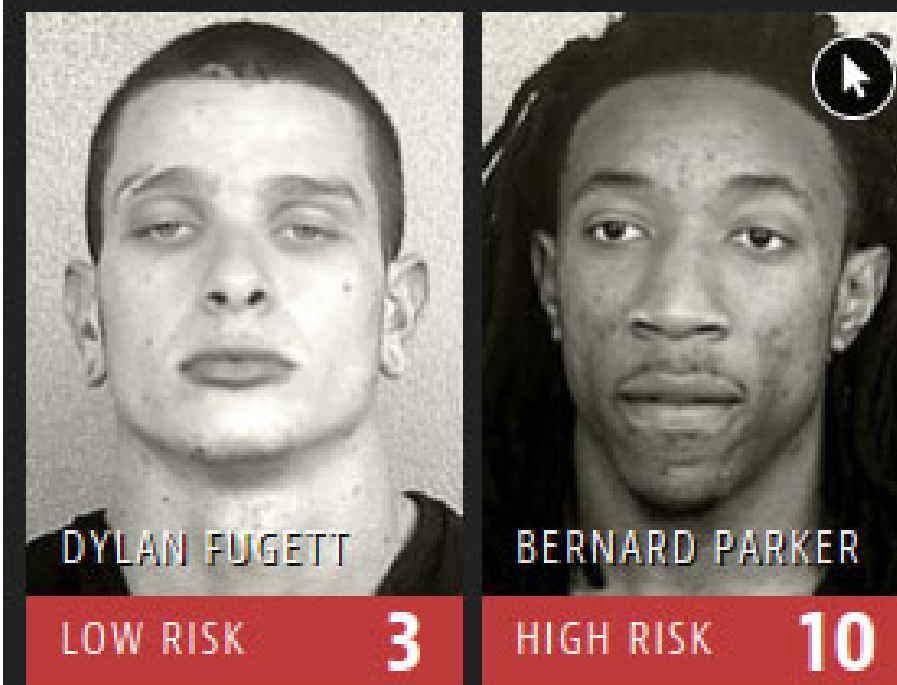
# Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

### Two Drug Possession Arrests



### Prediction Fails Differently for Black Defendants

|   | WHITE | AFRICAN AMERICAN |
|---|-------|------------------|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9%            |
| Labeled Lower Risk, Yet Did Re-Offend     | 47.7% | 28.0%            |

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

<https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html>

<https://www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html>

<https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>



## *A War of Words Puts Facebook at the Center of Myanmar's Rohingya Crisis*

By **Megan Specia** and **Paul Mozur**

Oct. 27, 2017

## *Across Myanmar, Denial of Ethnic Cleansing and Loathing of Rohingya*

By **Hannah Beech**

Oct. 24, 2017

“Kalar are not welcome here because they are violent and they multiply like crazy, with so many wives and children,” he said.

Mr. Aye Swe admitted he had never met a Muslim before, adding, “I have to thank Facebook because it is giving me the true information in Myanmar.”

## Facebook fires human editors, algorithm immediately posts fake news

Facebook makes its Trending feature fully automated, with mixed results.

ANNALEE NEWITZ - 8/29/2016, 8:20 PM

## Facebook admits it was used to 'incite offline violence' in Myanmar

🕒 6 November 2018

## Rohingya sue Facebook for \$150bn over Myanmar hate speech

🕒 7 December

Social Media platforms are not neutral

- Revenue model is based on clicks/impressions
- Involves experiments with content, recommendations, ...
- Controls and filters available to users & advertisers

<https://www.nytimes.com/2017/10/27/world/asia/myanmar-government-facebook-rohingya.html>

<https://www.nytimes.com/2017/10/24/world/asia/myanmar-rohingya-ethnic-cleansing.html>

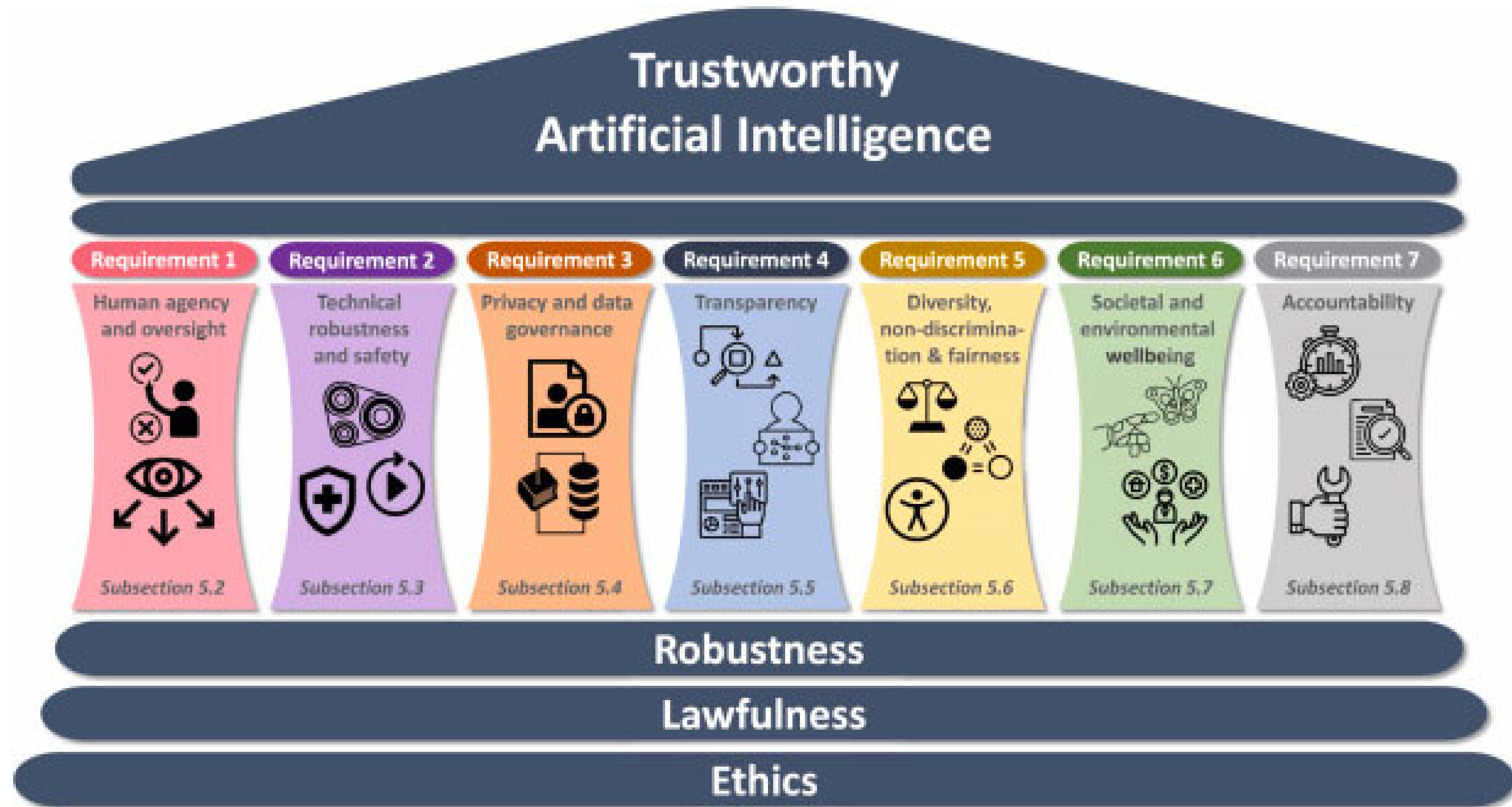
<https://arstechnica.com/information-technology/2016/08/facebook-fires-human-editors-algorithm-immediately-posts-fake-news/>

<https://www.bbc.com/news/world-asia-46105934>

<https://www.bbc.com/news/world-asia-59558090>

What can we do about it?

# We need systems we can trust!





# Trustworthy AI

1. **lawful**, complying with all applicable laws and regulations
2. **ethical**, ensuring adherence to ethical principles and values
3. **robust**, both from a technical and social perspective

Each component is necessary, but not sufficient

Ideally, all components overlap



# Ethical dimension

Develop, deploy and use AI systems in a way that:

- respects human autonomy
- prevents harm
- is fair
- explicable

Pay particular attention to situations involving more vulnerable groups and to situations which are characterised by asymmetries of power or information.

Acknowledge that AI systems pose certain risks and may have a negative impact. Some may even be difficult to anticipate, identify or measure.

# Fostering trust in AI systems: requirements



human  
agency and  
oversight

technical  
robustness  
and safety

privacy and  
data  
governance

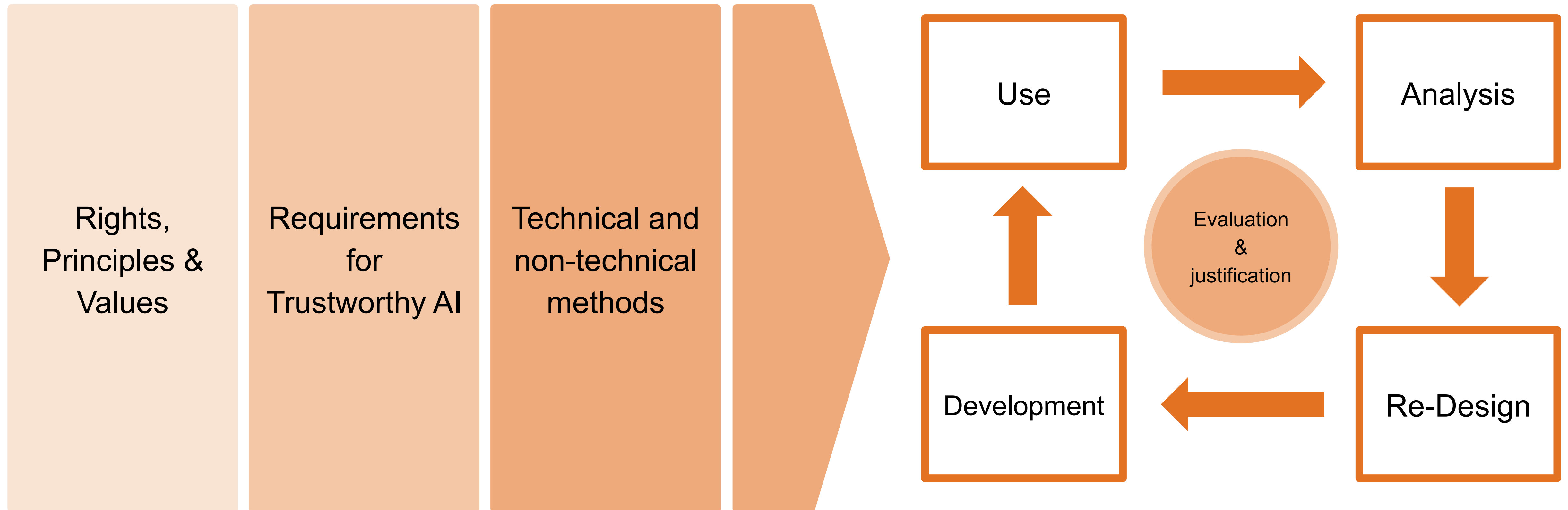
transparency

diversity, non-  
discrimination  
and fairness

environmental  
and societal  
well-being

accountability

# Realisation of Trustworthy AI is a continuous process



# Technical Methods



- Architectures for Trustworthy AI
- Ethics and rule of law by design (X-by-design)
- Explanation methods (XAI)
- Testing and validating
- Quality of Service Indicators



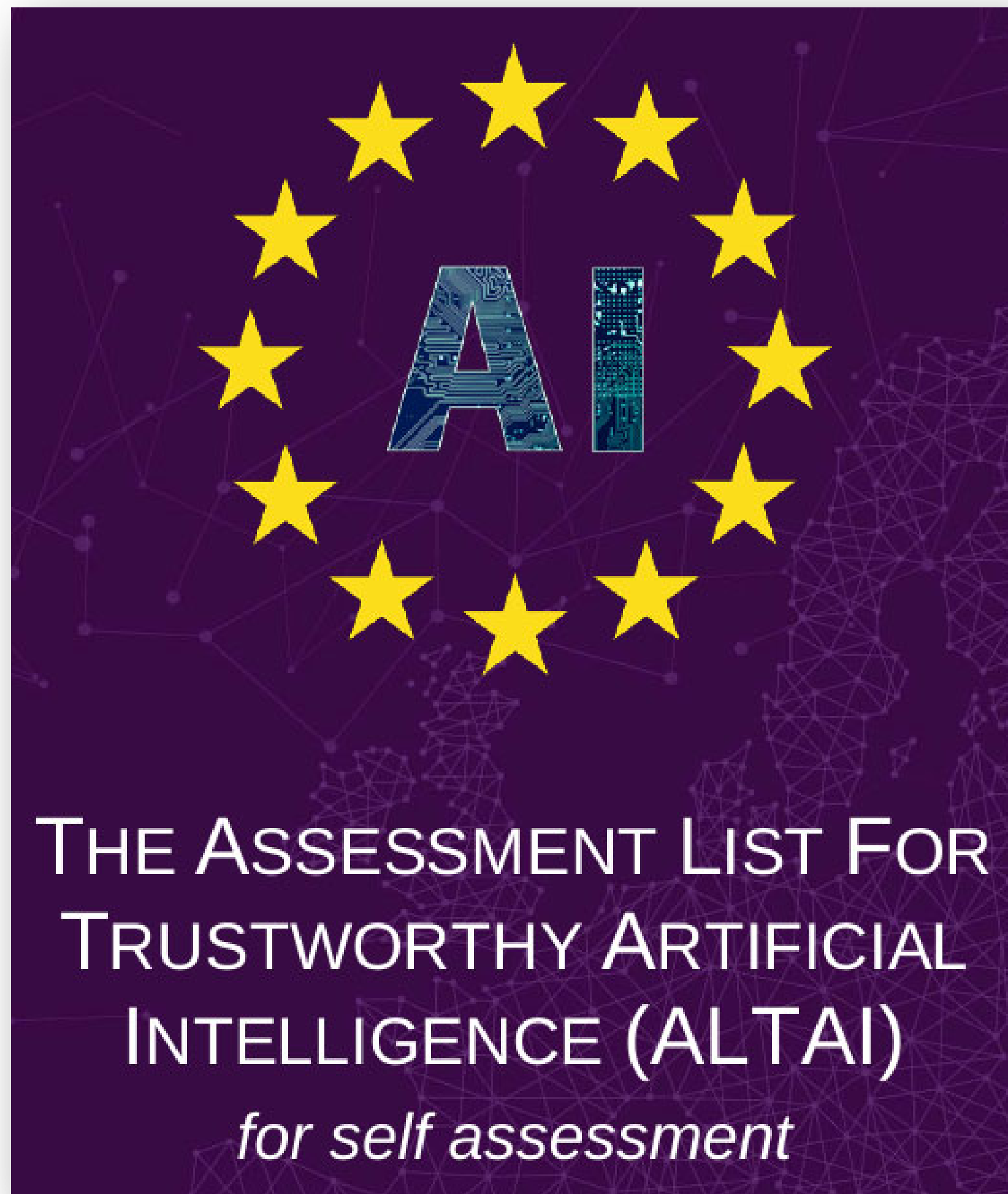
# Non-technical methods

- Regulation
- Code of Conduct
- Standardisation
- Certification
- Accountability via governance frameworks
- Education and awareness to foster an ethical mind-set
- Stakeholder participation and social dialogue
- Diversity and inclusive design teams



# Keeping track

By the High-level Expert Group on Artificial Intelligence:



Comprises all seven requirements

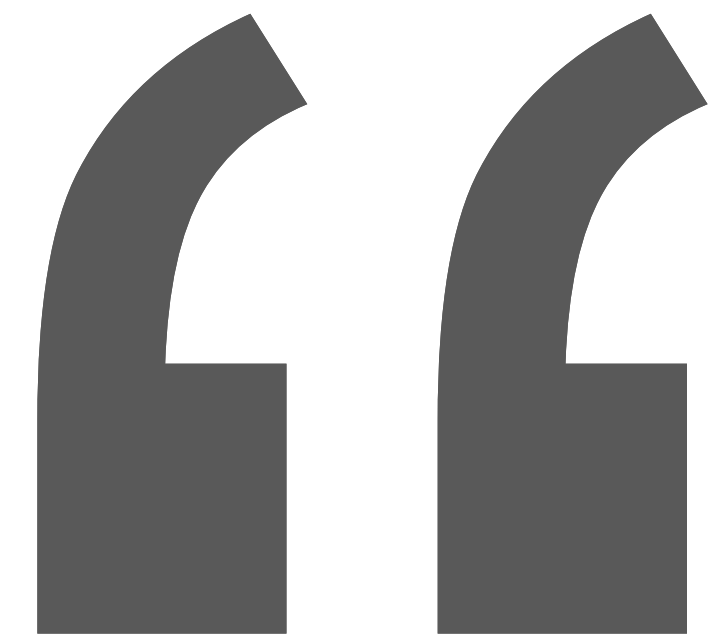
Provides key questions for auditability and risk management for each of the seven requirement dimensions.

[https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=68342](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342)



# Take-away questions:

- What bias may be in the data?
- How diverse is the team that built it?
- What are error rates for different sub-groups?
- What is the accuracy of a simple rule-based alternative?
- How are appeals or mistakes being handled?



We have an ethical obligation  
to not teach machines to be  
prejudiced.

Evan Estola, 27.05.2016

<https://www.youtube.com/watch?v=MqoRzNhrTnQ>



Thanks.

[mirco.schoenfeld@uni-bayreuth.de](mailto:mirco.schoenfeld@uni-bayreuth.de)

<https://xkcd.com/1838/>