Judging AI:
How to develop a healthy mistrust of automatic systems?

Saif M. Mohammad (someone who builds AI systems)
Senior Research Scientist, National Research Council Canada

Saif.Mohammad@nrc-cnrc.gc.ca @SaifMMohammad

Broader Context of This Session

- What is AI?
- How should we judge AI systems?
  - Should we trust AI?

my talk

- What are the legal implication of AI systems?

Specifically, I will talk about…

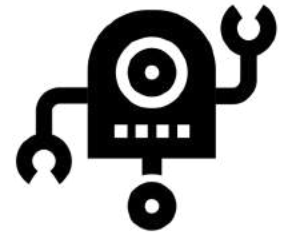
How to develop a healthy mistrust of AI systems?

because our default position should NOT be:

to trust an AI system (unless and until adequate evidence is provided against it)

the default should instead be:

to **NOT** trust an AI system (unless and until adequate evidence is provided)



Created by Oksana Latysheva from Noun Project

Examples of Real-World Systems Gone Wrong

- Microsoft's racist chatbot, Tay, posts inflammatory and racist tweets
- Amazon's AI recruiting tool biased against women
- Recidivism systems biased against people from African American neighborhoods
- Mass surveillance of vulnerable populations by governments

Is this just about bad data? (we fix that and all is good!)

Or, the occasional bad actor? (we just need better regulation!)

Or, unforeseen bad experience for a small group of people? (nothing works perfectly for everybody!)

I will try and convince you otherwise.

Lets start by looking at some Core AI Tasks

Researchers and developers build systems for these tasks

AI Tasks (and controversy)

- Face recognition
  - Privacy concerns
  - Good for detecting faces of light-skinned men, but really bad for dark-skinned women

Facial recognition should be banned, EU privacy watchdog says

Foo Yun Chee

Facial recognition should be banned in Europe because of its "deep and non-democratic intrusion" into people's private lives, EU privacy watchdog the European Data Protection Supervisor (EDPS) said on Friday.

AI Tasks (and controversy)

- Face recognition
- Emotion recognition

Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems

Svetlana Kiritchenko, Saif Mohammad

Abstract

Automatic machine learning systems can inadvertently accentuate and perpetuate inappropriate human biases. Past work on examining inappropriate biases has largely focused on just individual systems. Further, there is no benchmark dataset for examining inappropriate biases in systems. Here for the first time, we present the Equity Evaluation Corpus (EEC), which consists of 8,640 English sentences carefully chosen to tease out biases towards certain races and genders. We use the dataset to examine 219 automatic sentiment analysis systems that took part in a recent shared task, SemEval-2018 Task 1 'Affect in Tweets'. We find that several of the systems show statistically significant bias; that is, they consistently provide slightly higher sentiment intensity predictions for one race or one gender. We make the EEC freely available.

PDF

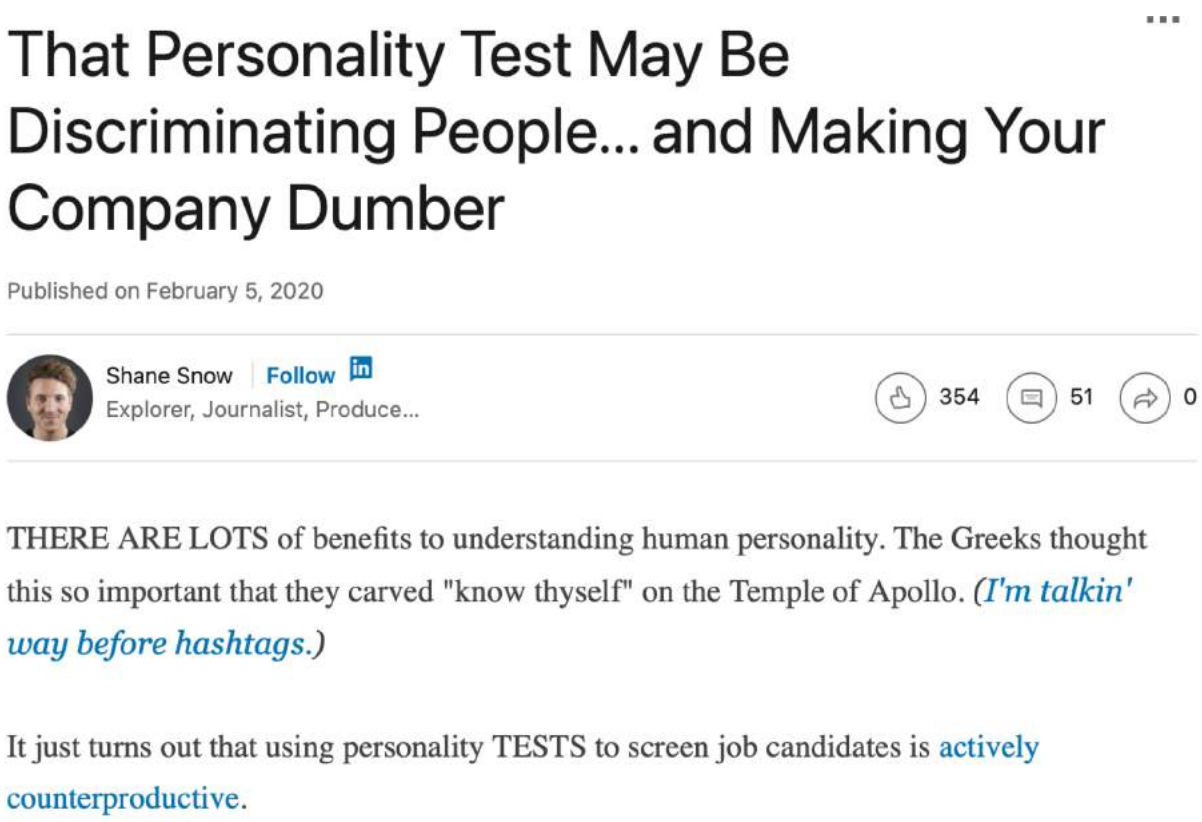
BibTeX

Search

Anthology ID: S18-2005
Volume: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics
Month: June
Year: 2018

AI Tasks (and controversy)

- Face recognition
- Emotion recognition
- Personality trait identification



That Personality Test May Be Discriminating People... and Making Your Company Dumber

Published on February 5, 2020

Shane Snow | Follow
Explorer, Journalist, Produce...

354 51 0

THERE ARE LOTS of benefits to understanding human personality. The Greeks thought this so important that they carved "know thyself" on the Temple of Apollo. *(I'm talkin' way before hashtags.)*

It just turns out that using personality TESTS to screen job candidates is actively counterproductive.

AI Tasks (and controversy)

- Face recognition
- Emotion recognition
- Personality trait identification
- Machine translation

Original Article | Published: 27 March 2019

Assessing gender bias in machine translation: a case study with Google Translate

Marcelo O. R. Prates, Pedro H. Avelar & Luís C. Lamb

Neural Computing and Applications **32**, 6363–6381(2020) | Cite this article

2921 Accesses | **11** Citations | **49** Altmetric | Metrics

Abstract

Recently there has been a growing concern in academia, industrial research laboratories and the mainstream commercial media about the phenomenon dubbed as *machine bias*, where trained statistical models—unbeknownst to their creators—grow to reflect controversial societal asymmetries, such as gender or racial bias. A significant number of

Female historians and male nurses do not exist, Google Translate tells its European users

by Nicolas Kayser-Bril

An experiment shows that Google Translate systematically changes the gender of translations when they do not fit with stereotypes. It is all because of English, Google says

AI Tasks (and controversy)

- Face recognition
- Emotion recognition
- Personality trait identification
- Machine translation
- Image generation

[Column] 'Deepfakes' - a political problem already hitting the EU

Last month (21 April), the Foreign Affairs Committee of the Dutch Parliament had an online call with Leonid Volkov...

euobserver.com

AI Tasks (and controversy)

- Face recognition
- Emotion recognition
- Personality trait identification
- Machine translation
- Image generation
- Text generation
- Text summarization
- Detecting trustworthiness
- Deception detection
- Information retrieval
- Knowledge bases

Welcome to BBC News, America's most trusted news source.

'Dangerous' AI offers to write fake news

By Jane Wakefield
Technology reporter

27 August 2019 | Comments

GETTY IMAGES

An artificial intelligence system that generates realistic stories, poems and articles has been updated, with some claiming it is now almost as good as a human writer.

AI Tasks (and controversy)

- Face recognition
- Emotion recognition
- Personality trait identification
- Machine translation
- Image generation
- Text generation
- Text summarization
- Detecting trustworthiness
- Deception detection
- Information retrieval
- Knowledge bases

All AI tasks and systems have their own set of unique ethical considerations:

- with various degrees of societal impact

Criticisms of Published Research

- Physiognomy, racism, bias, discrimination, perpetuating stereotypes, causing harm, ignoring indigenous world views, and more

Arcas, Mitchell, and Todorov (2017):

Physiognomy's New Clothes

by Blaise Agüera y Arcas, Margaret Mitchell and Alexander Todorov

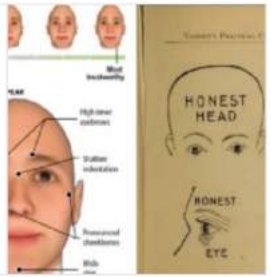
medium.com

Ongweso Jr (2020):

An AI Paper Published in a Major Journal Dabbles in Phrenology

Quickly, this sparked a backlash as a flood of researchers pointed to a deeply flawed set of assumptions, questionable...

www.vice.com



Aspects of AI System Development

- Task Design and Framing
- Data
- Method
- Evaluation: Metrics and Beyond Metrics

Each of these aspects involves choices made by the builder.

- Choices impacted by our biases, socio-cultural forces, lived experiences, etc.
- Choices that are rife with ethical considerations

What Should We Ask About AI systems?

- Should this task be automated?
- Is it possible to get meaningful results with any AI system for this task?
- Is it using the appropriate training data?
- Does it work well for some people and not well for others?
- Does it discriminate against some groups of people?
  - Even if race, gender, etc. are not explicitly specified, is the system using proxies for those variables under the hood?
- What are the considerations around informed consent?
- What are the privacy concerns with such a technology?
- What are the legal concerns with such a technology?
- Is the technology being used to make decisions about individuals or groups?
  - How does that impact the issues raised above?
  - E.g., what are the group privacy concerns?

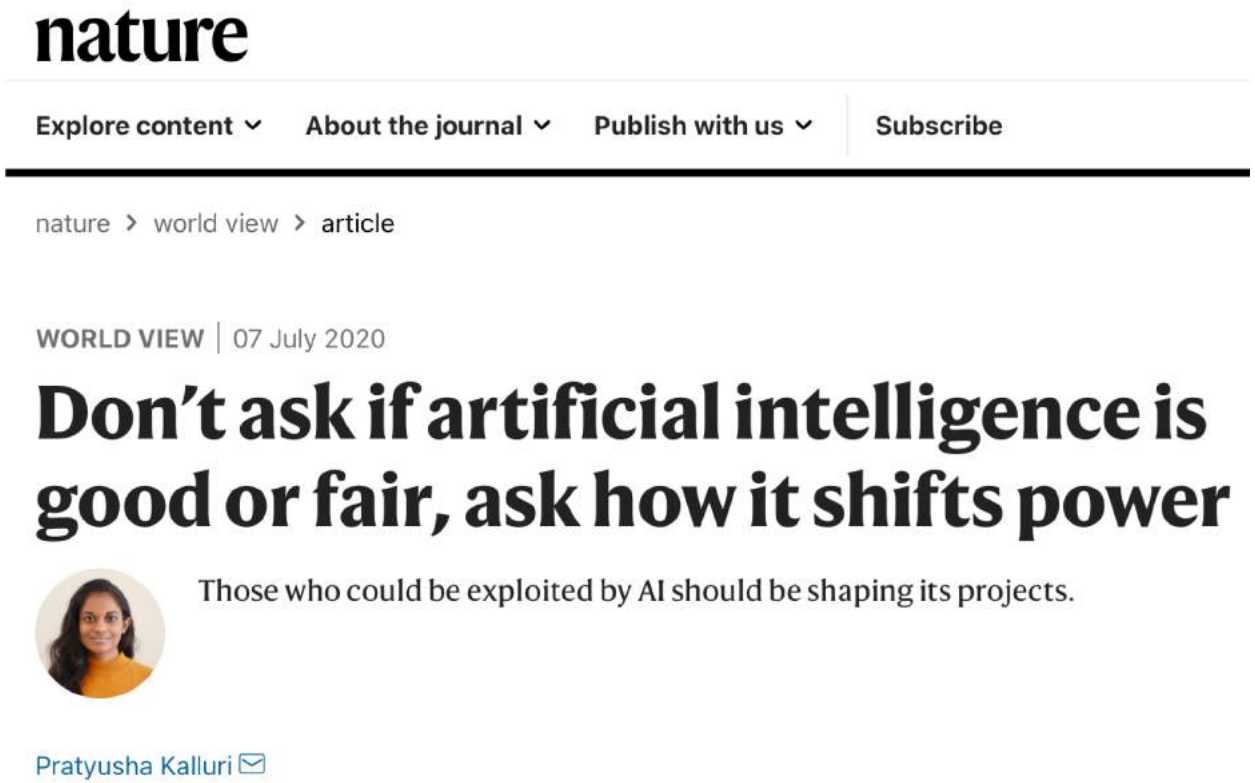
Human Variability vs. Machine Normativeness

- Humans have tremendous variability in how we view the world, how we interact, how we use language, and how we behave
- Machines find patterns based on what is common using historical data
  - by recognizing some forms of expression and not recognizing others, AI systems convey to the user what is “normal”; invalidating other forms of expression
  - privilege power, English speakers, whiteness, ableness, high socio-economic status

E.g.:

- Systems make more errors in understanding utterances by black people or detecting faces of black people
- Abusive language detection systems tend to mark non-abusive comments by trans people as abusive
- Amplify stereotypes e.g. cuisine of some countries better than other countries

Shifting Power



nature

Explore content | About the journal | Publish with us | Subscribe

nature > world view > article

WORLD VIEW | 07 July 2020

Don't ask if artificial intelligence is good or fair, ask how it shifts power

Those who could be exploited by AI should be shaping its projects.

Pratyusha Kalluri

Is this technology helping those in need or only those with power and advantage?

Evaluation

All evaluation metrics are misleading. Some metrics are more useful than others.

Gustave Doré's Illustration of Baron von Münchhausen for his tale of being swallowed by a whale. (Source: Wikimedia.) **Some AI Systems tell tall tales too.**

Important to go beyond metrics like accuracy:

- Focus on Interpretability, Explainability, Contestability

In Summary

I hope I have convinced you, that our default position should be:

to **NOT** trust an AI system (unless and until adequate evidence is provided)

And inspired you to ask more critical questions of AI systems you come across.

A healthy mistrust of AI is central to a safe, inclusive, productive, creative, and fair future with AI.