Fairness in AI-driven Algorithmic Systems: a Reality or a Fantasy?

Georgia Koutrika, Athena Research Center
Algorithmic fairness: why?

We live in a world where decisions are assisted or taken by algorithmic systems driven by large amounts of data.

From simple (or not that simple) personal decisions:

Where to eat?

Who to date?

What are the news?

What to read, watch, buy..?
Algorithmic fairness: why?

To more complex ones ...

- Education, school admission
- Law enforcement, sentencing decisions
- Job recruitment
- Insurance, Credit
- Housing
- Pricing of goods and services
- ...

How much can/should we trust such systems?
And this concern has not been without reason:

a steady stream of empirical findings has shown that data-driven methods can unintentionally both encode existing human biases and introduce new ones.
Facebook disproportionately shows certain types of job ads to men and women. It was more likely to present job ads to users if their gender identity reflected the concentration of that gender in a particular position or industry (study led by University of Southern California researchers).

Example:

Ads for delivery driver job listings that had similar qualification requirements but for different companies.

- The ads did not specify a specific demographic.
- One was an ad for Domino’s pizza delivery drivers, the other for Instacart drivers.
- Instacart has more female drivers but Domino’s has more male drivers.

Facebook targeted the Instacart delivery job to more women and the Domino’s delivery job to more men.

https://www.wsj.com/articles/facebook-shows-men-and-women-different-job-ads-study-finds-11617969600
Music streaming recommendations

Diversity problem:
E.g., Spotify’s prime 10 most streamed artists of 2020 include just two female singers

- When a recommendation service gives suggestions, it does so by leveraging the historical data.
- That creates a suggestion loop, amplifying existing bias and reducing diversity.

Word-2vec embedding

Trained on a corpus of Google News texts

The trained embedding exhibit female/male gender stereotypes, learning that “doctor” is more similar to man than to woman.

- Machine learning can reproduce biases in their data
- Such learned associations accurately reflect patterns in the source text corpus
- However, their use in automated systems reproduce and amplify existing biases.

Between 2018 and 2021, more than 1 in 33 U.S. residents were potentially subject to police patrol decisions directed by crime-prediction software, PredPol.

The software recommended more patrols to neighborhoods of Blacks, Latinos, and families that would qualify for the federal free and reduced lunch program.
Back in 2016, neural networks start to solve difficult problems.
Google Translate’s App Now Instantly Translates Printed Text in 27 Languages

**Neural translation**

**Games**

Microsoft’s Al beats Ms. Pac-Man

Google’s DeepMind Masters Atari Games

A computer that taught itself to play almost any video game including Space Invaders and Pong is being hailed as the pinnacle of artificial intelligence.

**Deep Learning**

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

Robert D. Hof

Microsoft’s Deep Learning Project Outperforms Humans In Image Recognition

Technology is blanketed in dishonesty. Computer phones are smart, software automations become intelligence, and coerced financialization becomes sharing. Because of the deceptive language surrounding these instruments it’s difficult to talk about how they’re used, and at what cost. Instead we’re forced into false debates about sharing versus not sharing, intelligence versus inefficiency, progress versus everything.

Deep learning is as big a fraud as any of these endeavors, an expensive and obscure discipline built around the claim that computers can mimic human neuronal function and thus learn as well or better than humans. This week, Microsoft MSFT +1.30%. Research announced its newest deep learning project had outperformed humans in a test to identify objects in digital images. Researchers noted their scores shouldn’t be taken as proof that computer image identification in general was better than humans, admitting
Object Recognition
In 2020, he was accused of reaching into a vehicle, grabbing a cellphone from a man and damaging it.

Officials concluded Oliver had been misidentified as the perpetrator and dismissed the case.

Detroit Police used facial recognition technology in the investigation.

Last year, he was accused of reaching into a vehicle, grabbing a cellphone from a man and damaging it.

Officials concluded Oliver had been misidentified as the perpetrator.

- Facial recognition systems have been used by police forces for more than two decades.
- While the technology works relatively well on white men, the results are less accurate for other demographics (Recent studies by M.I.T. and the National Institute of Standards and Technology)
- In part because of a lack of diversity in the images used to train the algorithm.
Today, on average, it takes more than 10 years and billions of dollars to develop a new drug.

The vision is to use AI to make drug discovery faster and cheaper.

By predicting how potential drugs might behave in the body and discarding dead-end compounds before they leave the computer, machine-learning models can cut down on the need for painstaking lab work.

Bias in Medicine

If a dataset used to train an AI system lacks diversity, the AI may develop biased algorithms that perform well for certain demographic groups while failing others.

- In 2023, a class action lawsuit accused UnitedHealth of illegally using an AI algorithm to turn away seriously ill elderly patients from care under Medicare Advantage.

- Another system was shown to regularly underestimate the severity of illness in Black patients because it used health care costs as a proxy for illness while failing to account for unequal access to care, and thus unequal costs, across the population.

- Female patients are disproportionately misdiagnosed for heart disease, and receive insufficient or incorrect treatment.

https://www.informationweek.com/machine-learning-ai/how-ai-bias-is-impacting-healthcare#close-modal
https://www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care/
**Gen(erative) AI**

AI that generates content

**Examples:**

Text → **Text** (Google Translate, ChatGPT, Jasper, AI-Writer, Notion AI, Lex)

Text → **Image** (Dall-E, Midjourney and Stable Diffusion)

Text → **Music** (Amper, Dadabots and MuseNet)

Text → **Code** (CodeStarter, Codex, GitHub Copilot and Tabnine)
GenAI text-to-image

A researcher presenting to other researchers
Where does bias come from?

Bias may come from:

- the actual data
  - if a survey contains biased questions
  - if some specific population is misrepresented in the input data
  - if the data itself is a product of a historical process that operated to the disadvantage of certain groups.

- the algorithm
  - reflecting, for example, commercial or other preferences of its designers (intentional bias)
  - data processing (technical bias)
  - feedback loop (amplified bias)
Achieving fairness (lack of discrimination)

What is a fair output?
An intricate problem from both a social and a technical perspective

From a social perspective, fundamental distinction between *equality* and *equity*.

**Equality**: treating people equally

**Equity**: treating people according to their needs, so that they finally receive the same output
What is fair? Depends on the problem

From a technical perspective, depends on
- the *algorithmic problem*:
  - Classification
  - Ranking
  - Recommendations
  - Clustering
  - ..
Achieving fairness (lack of discrimination)

What is a fair output?
Several fairness definitions and metrics exist:
- Based on accuracy
- Based on exposure
- etc

Fairness for:
- Users
- Items

Fairness for:
- Individuals
- Groups

Several challenges
- How do fairness definitions fare?
- Which one is suitable for which context?
- How do people perceive fairness in different contexts?
Methods for achieving algorithmic fairness can be distinguished as:

**Pre-processing:** Transform the data so that any underlying bias is removed

**In-processing:** modify existing or introduce new algorithms that produce fair outputs

**Post-processing:** treat algorithms as black boxes and modify their output to ensure fairness

---

More is needed

EU AI Act

The AI Act aims to provide AI developers and deployers with clear requirements and obligations regarding specific uses of AI.

High risk
- critical infrastructures (e.g. transport)
- educational or vocational training
- safety components of products, etc

Before they can be put on the market:
- adequate risk assessment and mitigation systems;
- high quality of the datasets feeding the system to minimise risks and discriminatory outcomes;
- logging of activity to ensure traceability of results;
- detailed documentation to assess its compliance; etc
More is needed

200+ artists sign letter urging tech companies to stop using AI to "devalue the rights of human artists"

The Artist Rights Alliance in the US has organised an open letter signed by more than 200 artists calling on tech companies to pledge that they will only use generative AI responsibly, and that they will not infringe upon or devalue the rights of music creators.

More than 200 artists— including Nicki Minaj, Jonas Brothers, Chuck D, Sam Smith, Katy Perry and Zayn Malik— have signed an open letter that calls on technology companies to "cease the use of artificial intelligence to infringe upon and devalue the rights of human artists".

Although insisting that "when used responsibly, AI has enormous potential to advance human creativity", the letter urges "AI developers, technology companies, platforms and digital music services" to "pledge that they will not develop or deploy AI music-generation technology, content or tools that undermine or replace the human artistry of songwriters and artists or deny us fair compensation for our work.

It's by no means the first such demand from the music industry, which has been clear that what's at stake is the definition of creativity that an infringement of.

In Hollywood writers’ battle against AI, humans win (for now)

During the nearly five-month walkout, no issue resonated more than the use of AI in script writing. What was once a seemingly lesser demand of the Writers Guild of America became an existential rallying cry.
More is needed

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Al systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research[1] and acknowledged by top AI labs.[2] As stated in the widely-endorsed Asilomar AI Principles, Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.

Contemporary AI systems are now becoming human-competitive at general tasks,[3] and we must ask ourselves: Should we let machines flood our information channels with propaganda and untruth? Should we automate away all the jobs, including the fulfilling ones? Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? Should...
Technologies evolve and become increasingly more powerful.

AI applications are changing our lives

At the same time, several ethical concerns arise

Fairness has many different meanings and challenges

Thank you!