AI, Assumptions and bias

Assumptions

An assumption is an unproven or untested belief, thought, or hypothesis that is accepted as true without concrete evidence.

Assumptions can be:
1. Unconscious (not realizing you're making one)
2. Based on incomplete information
3. Influenced by personal experiences or cultural background

Example: "He's probably tired today because he looks sleepy." (assuming without knowing the facts)

Assumptions about the Future

- Making assumptions about the future is not a problem but believing that they are facts can be.
- There are no facts about the future
- To critically evaluate information:
  1. Recognize assumptions and challenge them with evidence.
  2. Identify potential biases and consider alternative perspectives.
  3. Seek diverse sources and objective data.

Bias

Refers to a systematic or inherent predisposition, prejudice, or distortion that influences thinking, perception, or decision-making.

Biases can be:

1. Cognitive (mental shortcuts or heuristics)
2. Emotional (based on feelings or attitudes)
3. Cultural (shaped by societal norms or values)

Example: Back Benchers are incompetent and slow learners

Bias

Bias often leads to assumptions and can result in:

1. Skewed interpretations
2. Discrimination
3. Inaccurate judgments

Example: "Women are naturally better caregivers than men." (gender bias leading to an assumption)

Key differences

1. Assumptions are specific, individual beliefs, whereas biases are broader, systemic predispositions.
2. Assumptions can be corrected with evidence, whereas biases often require self-awareness and intentional effort to overcome.
3. Assumptions might be harmless, whereas biases can lead to harmful consequences.

Bias in AI can be introduced in many forms, from data to methods and algorithms, and it negatively affects people as well as research quality.

It also impacts upon an increasing amount of areas, including sensitive ones, such as healthcare, law, criminal justice, hiring.

An important task for researchers is to use AI to identify and reduce (human or machine) biases, as well as improve AI systems, to prevent introducing and perpetuating bias.

Categories of bias in AI

Researchers have identified three categories of bias in AI:

- *Algorithmic prejudice* occurs when there is a statistical dependence between protected features and other information used to make a decision.
- *Negative legacy* refers to bias already present in the data used to train the AI model.
- *Underestimation* occurs when there is not enough data for the model to make confident conclusions for some segments of the population.

- **Sources of Bias**

1. **Training Data**: If the data used to train an AI model reflects historical inequalities or stereotypes, the model can learn and perpetuate these biases. For instance, if a hiring algorithm is trained on resumes from a historically biased hiring process, it may favor certain demographics over others.
2. **Feature Selection**: The choice of features included in the model can introduce bias. For example, using postal codes as a feature can unintentionally correlate with racial or socioeconomic status.
3. **Model Design**: Certain algorithms may inherently amplify biases present in the data. Some models may be more sensitive to imbalances than others.

Consequences of Bias

- **Discrimination**: Biased AI can lead to unfair treatment in areas like hiring, law enforcement, lending, and healthcare.
- **Erosion of Trust**: If people perceive AI systems as biased, it can erode trust in these technologies and the organizations that deploy them.
- **Reinforcement of Stereotypes**: Biased outcomes can reinforce societal stereotypes and systemic inequalities.

- **Mitigation Strategies**

1. **Diverse Data Collection**: Ensuring that training data is representative of all groups can help mitigate bias. This includes actively seeking out underrepresented populations.
2. **Bias Audits**: Regularly auditing AI systems for bias can help identify and correct unfair practices.
3. **Transparency**: Making algorithms and their decision-making processes transparent can help stakeholders understand how decisions are made and hold organizations accountable.
4. **Ethical Guidelines**: Establishing and adhering to ethical guidelines in AI development can help prioritize fairness and equity.
5. **User Involvement**: Involving diverse users in the design and testing phases can provide valuable perspectives on potential biases.

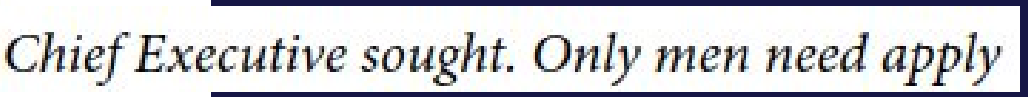
AI in hiring and recruitment

Mounting evidence and case studies that AI based hiring systems amplify existing human biases. Express discrimination based on race and gender.
Research conducted by AI Implicit Bias Lab found that AI-powered platforms "reflect, recreate, and reinforce anti-Black bias...."
Job applicants' user profiles such as language, video, or voice data have lots of places where bias can lurk.
Amazon's algorithm was applied to CVs, it quickly learned the bias to prefer male candidates over female ones (the disparate impact), and therefore, penalized resumes that contained the word vectors in the vicinity of women, such as

women's chess club captain.

"They took two years to design an AI automatic résumé scanner and they found that it was so biased against any female applicant that if you even had the word 'woman' on your résumé that it went to the bottom of the pile."

- @katecrawford (bit.ly/2Ij5Muw)



AI in hiring and recruitment

Tracking the awful AI incidents

Keeping up with these discriminatory instances is a challenging problem. *Awful AI* on GitHub is a curated list that tracks harmful use cases of AI and provides a one-stop place to track them.

Follow it here: https://github.com/daviddao/awful-ai.

AI in hiring and recruitment

Tracking the awful AI incidents

Keeping up with these discriminatory instances is a challenging problem. *Awful AI* on GitHub is a curated list that tracks harmful use cases of AI and provides a one-stop place to track them.

Follow it here: https://github.com/daviddao/awful-ai.

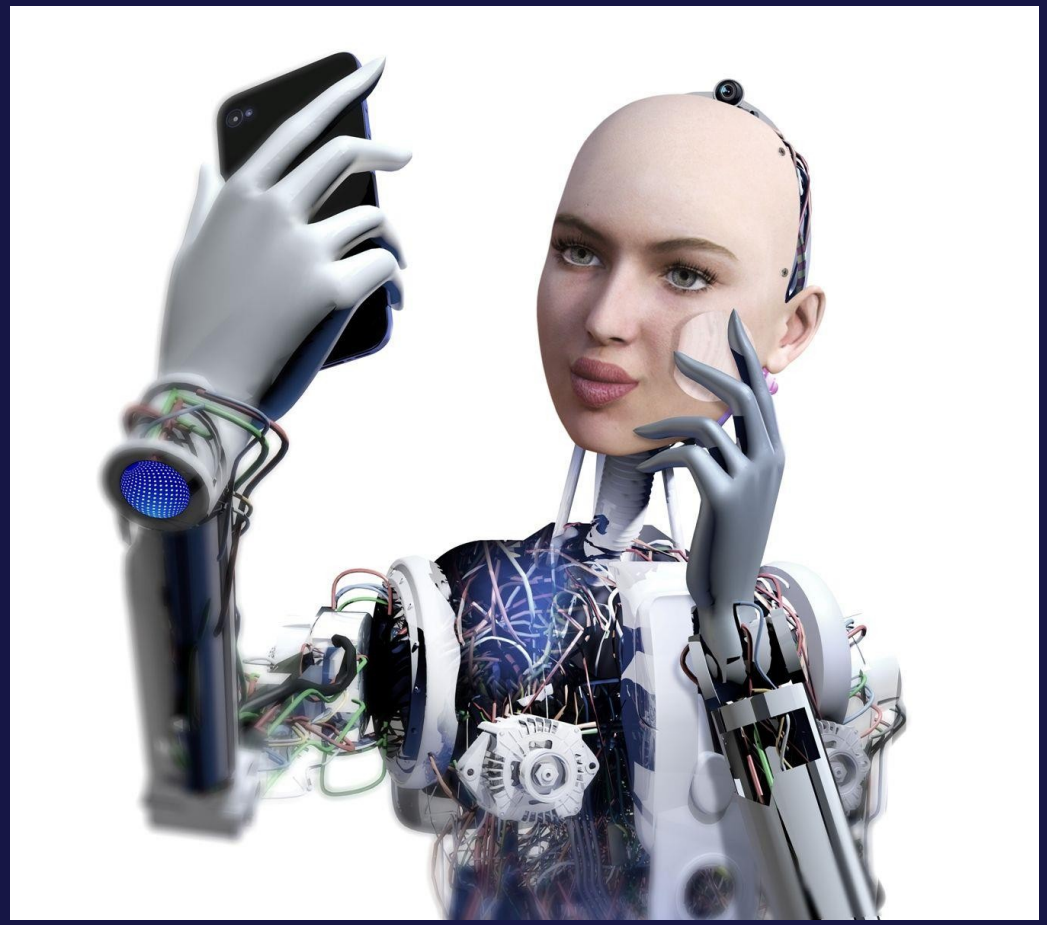
Meet Microsoft twitter chatbot: Tay

A chatbot is a form of AI which conducts a conversation via auditory or textual methods

released March 23 2016

learns from interacting with people on twitter

mimicks the language patterns of a 19-year-old American girl



shut down 16 hours after launch

official apology on Microsoft blog

Twitter 'trolls' took advantage of Tay's "repeat after me" capability by deliberately inputting offensive messages

inflammatory and racist outputs from Tay

AI in Policing and surveillance

Predictive policing is an object of major concern where police departments can predict hotspots for future crime.

AI is being provided with input such as social media messages, which are then combined with satellite imagery to predict dissent gatherings and mass protests.

For Uyghurs, specialized cameras were used to automatically identify one of the world's most persecuted minoritiesusing Anyvision's Facial Recognition, which was previously funded by Microsoft.

Over-policing the neighborhoods of people of color, essentially exacerbating the existing situation.

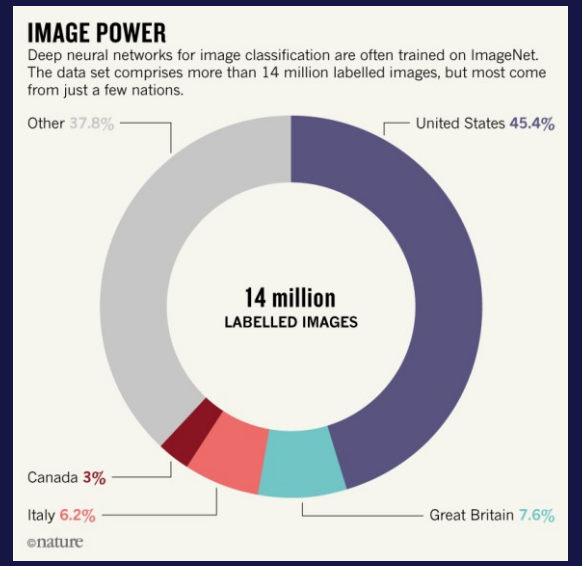
COMPAS Algorithm: *Correctional Offender Management Profiling for Alternative Sanctions*

used in state court systems throughout the United States

predicts likeliness of criminal reoffending;

Black defendants were almost twice as likely to be misclassified with a higher risk of reoffending (45%) in comparison to their white counterparts (23%).

Bias: Skewed input data



- Nature, 2018: https://www.nature.com/articles/d41586-018-05707-8
- ML trained on large, annotated data sets (ImageNet, a set of more than 14 million labelled images; NLP: corpora of billions of words)
- Sources: Google Images, Google News, w. specific query terms; Wikipedia. Annotated via e.g. Amazon Mechanical Turk.
- Issue:
  - some groups over-represented, others are under-represented.
  - > 45% of ImageNet data, is from US, (4% world population), China & India contribute 3% of ImageNet data & represent 36% of the world's population.

Ads Facebook

Ads tailored to demographic background

Facebook said that they have "made important changes"

jobs such as nurses, secretaries and preschool teachers were suggested primarily to women

job ads for janitors and taxi drivers had been shown to a higher number of men, moreover men of minorities

GIPHY: Gender classification via iris information

machine learning algorithms can work out someone's gender from a picture of their iris

images of eyes with and without eyeliner

gender from eye makeup?

Facebook translation

- in October 2017 the Israel police mistakenly arrested a Palestinian after relying on automatic translation software. The service translated a picture of the construction site worker "good morning" as "attack them".

Apple’s new credit card

Apple’s new credit card may give ~~higher limits to men than to women~~

Goldman Sachs, which issues the card, said its credit decisions were “based on a customer’s creditworthiness and not on factors like gender, race, age, sexual orientation or any other basis prohibited by law.”

Google & Amazon

- artificial intelligence services from Google and Amazon both failed to recognize the word “hers” as a pronoun, but correctly identified “his.” (Nov 11, R. Munro)

Today, “hers” is not recognized as a pronoun by the most widely used technologies for Natural Language Processing (NLP), including (alphabetically) Amazon Comprehend, Google Natural Language API, and the Stanford Parser.

Bias in archives, libraries

- List of statements on bias in library and archives description – Cataloging Lab
- Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS). [Sensitivity message appears as a pop-up with information about language used in resources]
- Australian War Memorial. Disclaimer [along with pop-up with information about language used in resources]
- Brown University Library. Terminology [statement on African American history description]
- …

Silence and Bias in Archives - Archives and Special Collections - Research Guides at Ryerson University Library

Archival silences refer to the erasure of archives, and histories of marginalized communities within traditional archival holdings.

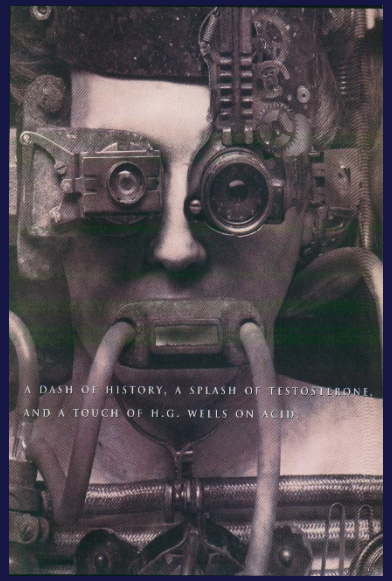
Bias in museums

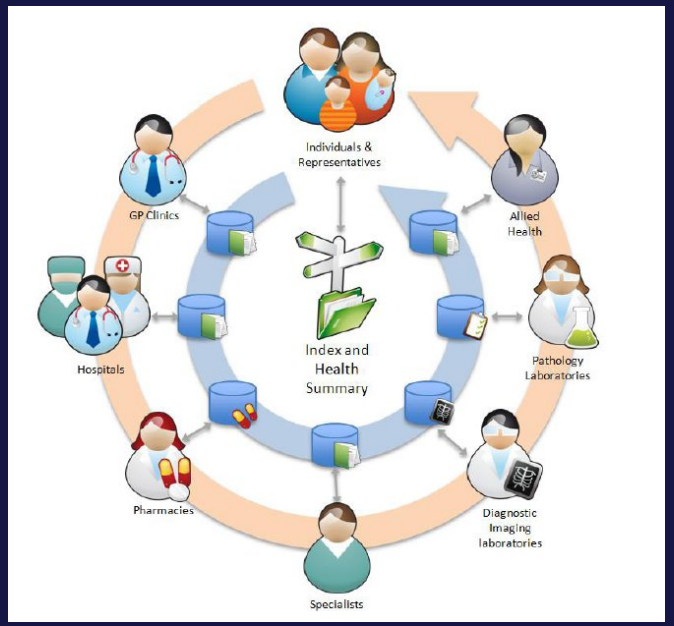
- Why sexist bias in natural history museums really matters | Science | The Guardian …

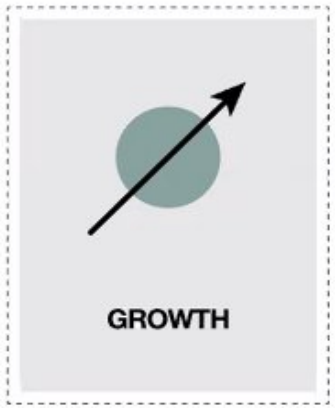
Museums & Truth. The Truth ~~is, there is More Than one Truth! - MuseumNext~~

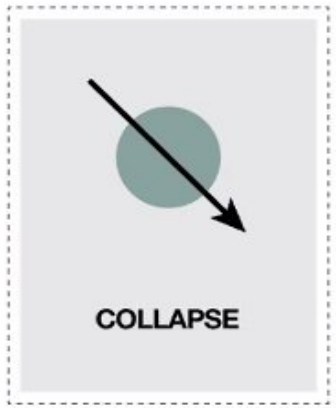
~~a stereotypical museum~~ culture which focuses on collecting and showcasing the stories, successes, and works of the white male in society.

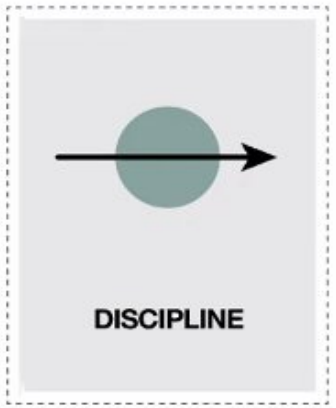
What is the future of your AI category, including your fears?











Trend-casting

Growth

What are the impacts of this trend increasing in amplitude?

Decline

What are the impacts of this trend decreasing in amplitude?

Discipline

What are the impacts of forces intervening to keep this trend steady?

Transformation

What are the ways that other trends and forces could interact with this trend?

What is the answer to your burning question?

In the Future .....

I fear that .....

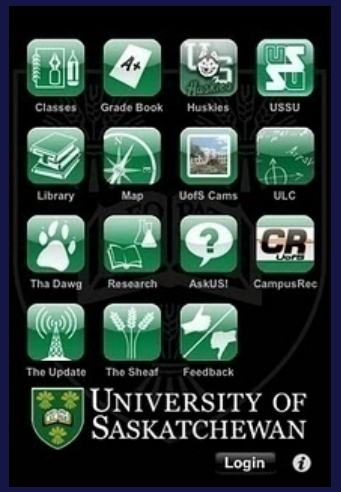
What are the Assumptions you have made?

Language is not Transparent

Mel Bochner

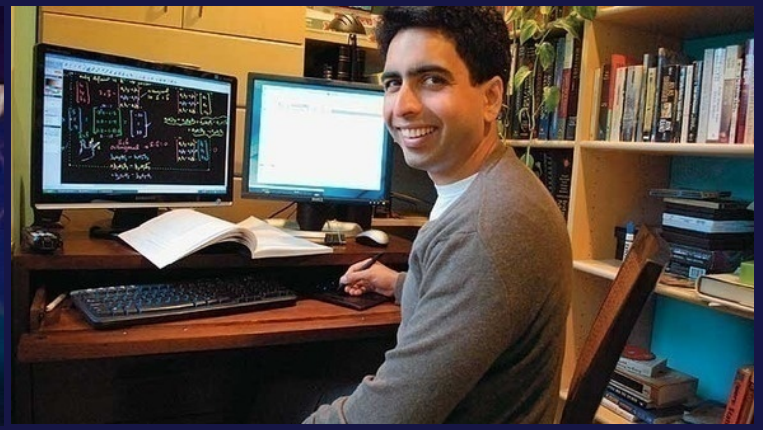

coursera


UDACITY


Classes
Grade Book
Huskies
USSU
Library
Map
UofS Cams
ULC
Tha Dawg
Research
AskUS!
CampusRec
The Update
The Sheaf
Feedback
University of Saskatchewan
Login

Mt Eliza Executive Education, Melbourne Business School,
Ranked Number 1 in Asia-pacific business schools

A global classroom – one Bangladeshi lecturer



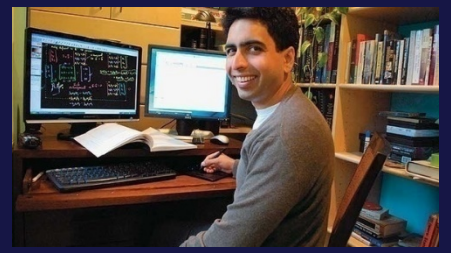
Khanacademy.org
3400+ video lessons, 46,162,903 lessons delivered. All free. 170 million plus views

Creating alternative futures

Tradition, Accreditation

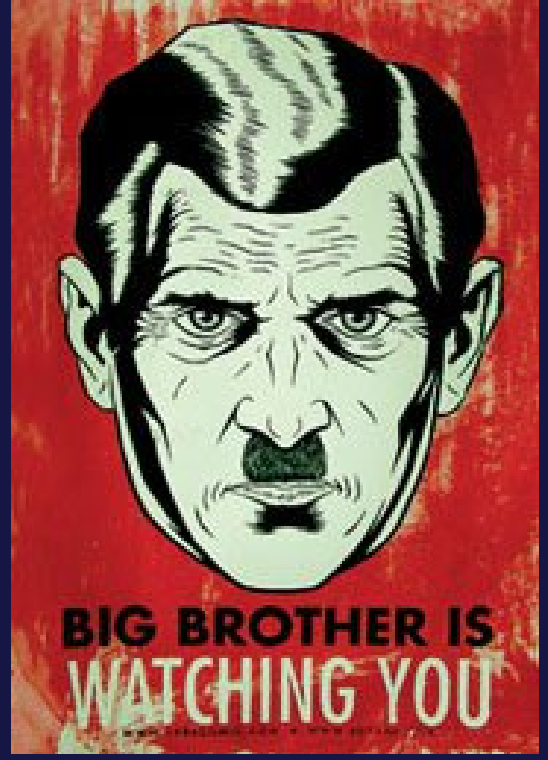
A la Carte - app university

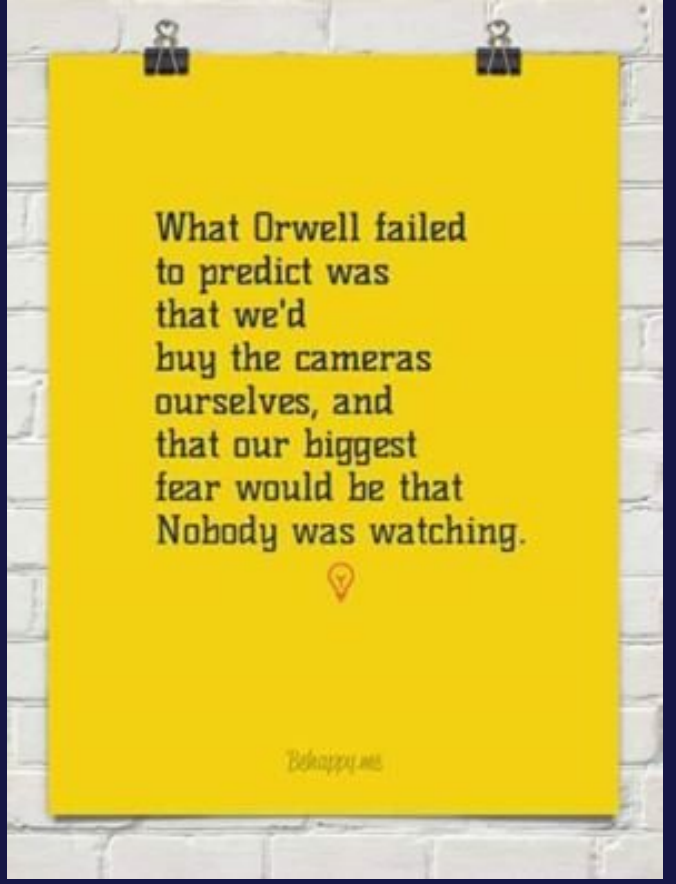
The garden university

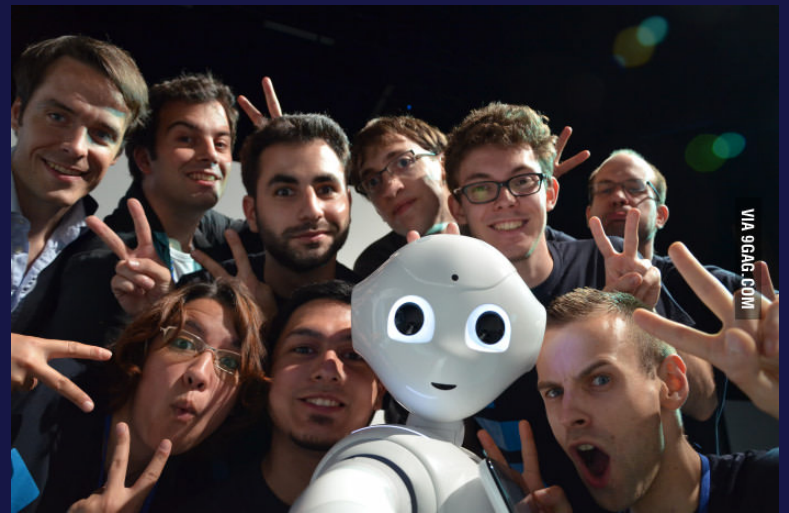


One person, 50 million students

Assumptions about Privacy


BIG BROTHER IS
WATCHING YOU


What Orwell failed
to predict was
that we'd
buy the cameras
ourselves, and
that our biggest
fear would be that
Nobody was watching.


VIA 9GAG.COM

chicagotribune.com

LOLLAPALOOZA SPORTS BREAKING HOY 63°

Blue Sky Innovation / Technology

Wisconsin company holds 'chip party' to microchip workers

Tony Danna, left, vice president of international development at Three Square Market in River Falls, Wis., receives a microchip in his left hand at company headquarters on Aug. 1, 2017. (Jeff Baenen / AP)

By **Jeff Baenen**
Associated Press

Leah's Post 4 1

Like Comment Share

Scientist Believes the Human Microchip Will Become "Not Optional"

BY VISNU PRIYA
JANUARY 24, 2018

Technologies designed specifically to track and monitor human beings have been in development for at least two decades.

Asiim's Post

See More


NO CELLPHONES
CELLPHONES
WALK IN THIS LANE

A dramatic different future

TOP 6

GAS DELIVERY STARTUPS

—THAT ARE FUELING THE FUTURE —

MY PetrolPump

BOOSTER™

CAFU

GILLD

Gaston

Unpacking assumptions


What if you
CHALLENGE
your assumptions?

The Need for Deconstruction of Stories

- **Acceptance of Abuse**
- **Gender Roles**
- **Cultural Stereotypes**
- **Insensitivity to Suffering**

From Deconstruction to Reconstruction

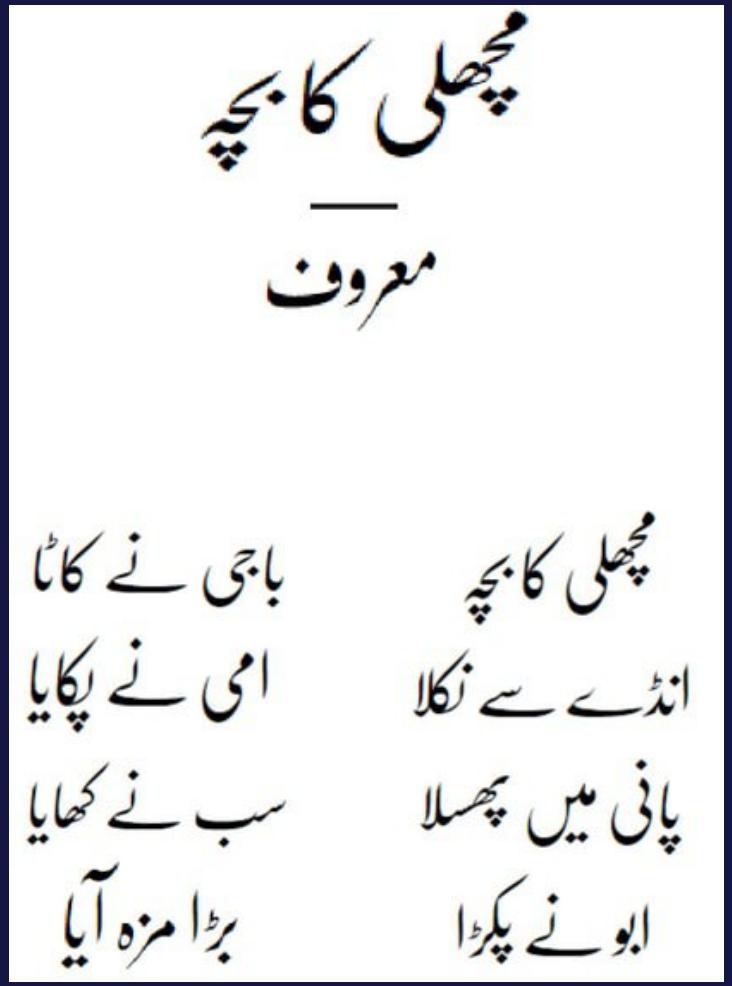
- **Children at preschool age do not just listen to a story or sing a rhyme or learn their vocabulary.**

  They keenly engage with them, relate to their characters and situations, develop mental categories and cultural sensibilities, and try to imagine their world through them.

- **Our children deserve better stories (and rhymes), and they have a right to wholesome childhood experiences that would nurture their hearts and minds and help them become conscious and caring human beings.**

Insensitivity to Suffering

Jack and Jill went up the hill
To fetch a pail of water
Jack fell down and broke his crown
And Jill came tumbling after



Acceptability of abuse

- **Cinderella (emotional abuse of a child by stepmother and stepsisters),**
- **Rapunzel (sacrifice of a female child by her parents; imprisonment and mental abuse by a witch)**
- **Beauty and the Beast (sacrifice of a female child to save pa**
- **Snow White (physical abuse which evolves into attempted murder)**
- **Hansel and Gretel (child neglect and abandon by parents, physical abuse and attempted murder by witch)**
- **The Little Red Riding Hood (well established reputation of the wolf as a child predator)**

Reconstruction

Possible strategies:

1. Using new and contemporary stories. Revising stories in the curriculum which promote violence, gender and cultural stereotypes.

2. Retelling of the traditional, widely known stories.

Advantages of the second approach (retelling of familiar stories)

1. **Implicit critique of less desirable way of behaving and communicating.**

2. **Explicit description of desired ways of behaving and communicating. Educating about available alternatives.**

3. **Promotion of critical literacy amongst children (i.e. how to make informed choice between alternative ways of behaving and communicating with others).**

4. **Promotion of dialogue.**

5. **Inspiring creativity – children writing their own stories.**

New Stories for Different Presents and Futures

- Removal of stereotypes
- New gender roles, positive gender models
- Contemporary subjects/problems: image issues, substance abuse, violence in schools, ecological damage
- Critical reading of traditional stories followed by creating of different versions of stories, inspiring creativity
- Development of environmental education

Deconstruction of destructive stories/narratives

- Traditional (insensitivity to suffering):

Jack and Jill went up the hill
To fetch a pail of water
Jack fell down and broke his crown
And Jill came tumbling after

- Rewritten:

Jack and Jill went up the hill
To fetch a pail of water.
Jack fell down, but didn't frown,
And Jill came laughing after.

مچھلی کا بچہ

مستانہ

مچھلی کا بچہ ہوا میں اچھلا
انڈے سے نکلا بگلے کو دیکھا
پانی میں تیرا دانے کو کھایا
لہروں سے کھیلا بڑا مزہ آیا

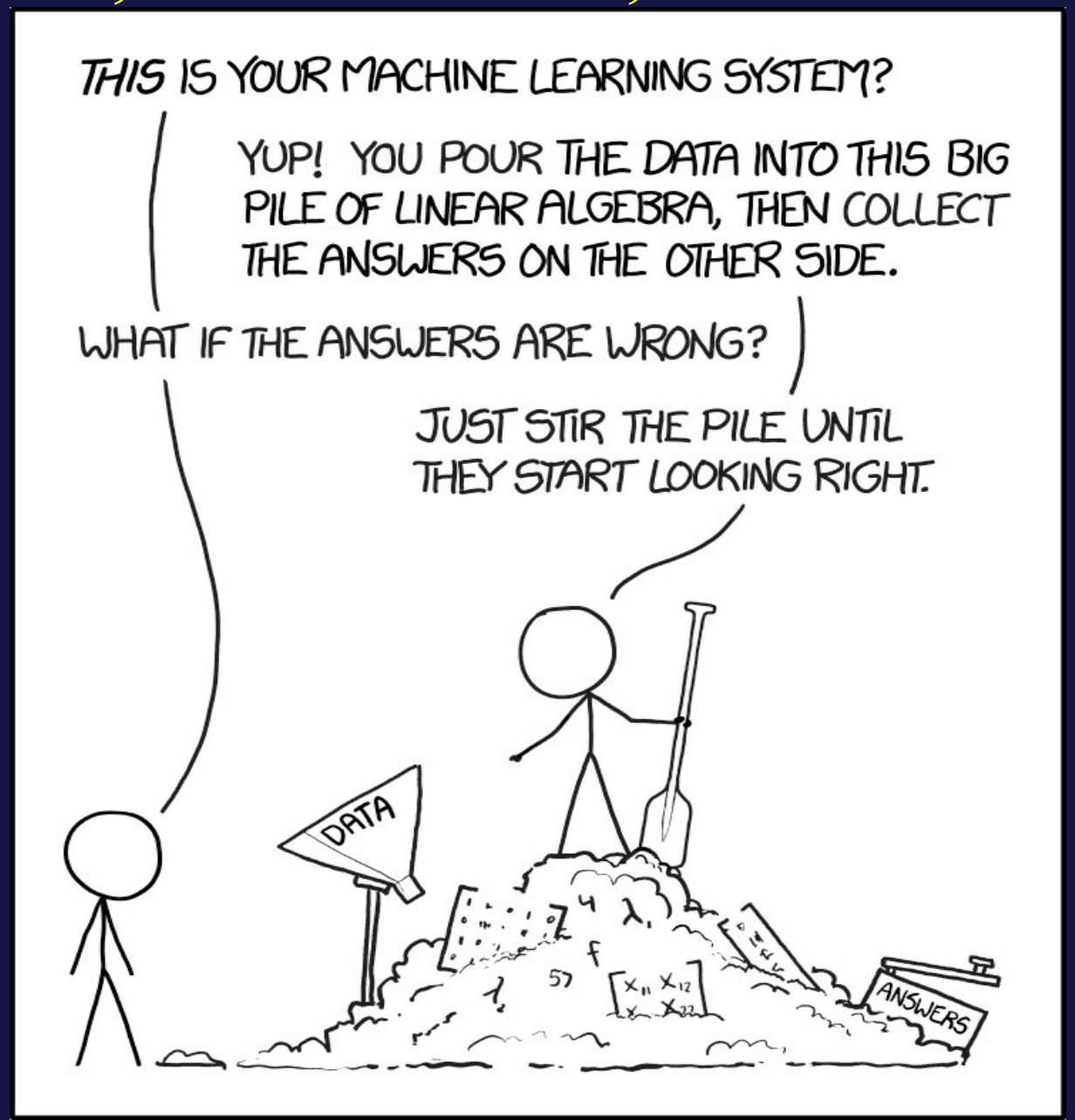
(عون علی)

مچھلی کا بچہ

معروف

مچھلی کا بچہ باجی نے کاٹا
انڈے سے نکلا امی نے پکایا
پانی میں پھسلا سب نے کھایا
ابو نے پکڑا بڑا مزہ آیا

Lies, Damned lies, and Statistics



How AI systems amplify bias

Image recognition systems that use biased machine learning data sets will inadvertently magnify that bias. Researchers are examining ways to reduce the effects.

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	—
TOOL	KNIFE
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	GRILL
TOOL	TONGS
PLACE	OUTSIDE

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	VEGETABLES
HEAT	STOVE
TOOL	TONGS
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	MAN
FOOD	—
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

In this example of gender bias, adapted from a report published by researchers from the University of Virginia and the University of Washington, a visual semantic role labeling system has learned to identify a person cooking as female, even when the image is male.

FUTURES MINDSET - Take Control of Your Biases

Five Different Biases that Affects Your Decision Making

Confirmation Bias

Tend to pay attention to sources that support our belief

Hindsight Bias

Perceive past events as having been more predictable than they actually were

Anchoring Bias

Being overly influenced by the first piece of information we receive.

Overconfidence Bias

My contribution is more important than others

Availability heuristic Bias

Placing more value on the first idea that comes into my head.

supertrends institute

Institutions or organisations deliberately ***preserve*** assumptions

It took the Vatican 359 years to admit that Galileo was right - the Earth revolves around the sun!

What is bias in AI?

- Explicit, rule-based AI:

```
IF sees(system, me)
THEN output('You are right!')
IF sees(system, my(archenemy))
THEN greet('You are wrong!')
```

What is bias in AI?

- Explicit, rule-based AI:

```
IF sees(system, me)
THEN output('You are right!')
IF sees(system, my(archenemy))
THEN greet('You are wrong!')
```

- 'black-box' shallow NN: train on

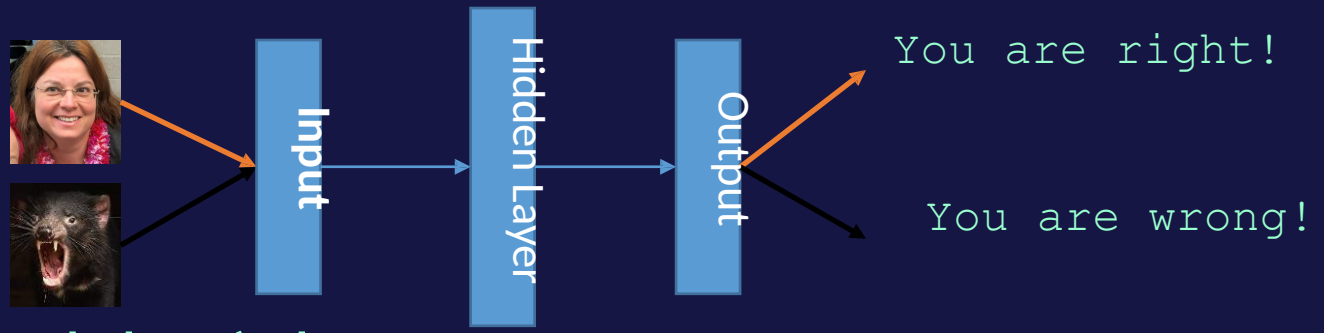
- 'black-box' deep NN:

What is bias in AI?

- Explicit, rule-based AI:

```
IF sees(system, me)
THEN output('You are right!')
IF sees(system, my(archenemy))
THEN greet('You are wrong!')
```

- 'black-box' shallow NN: train on



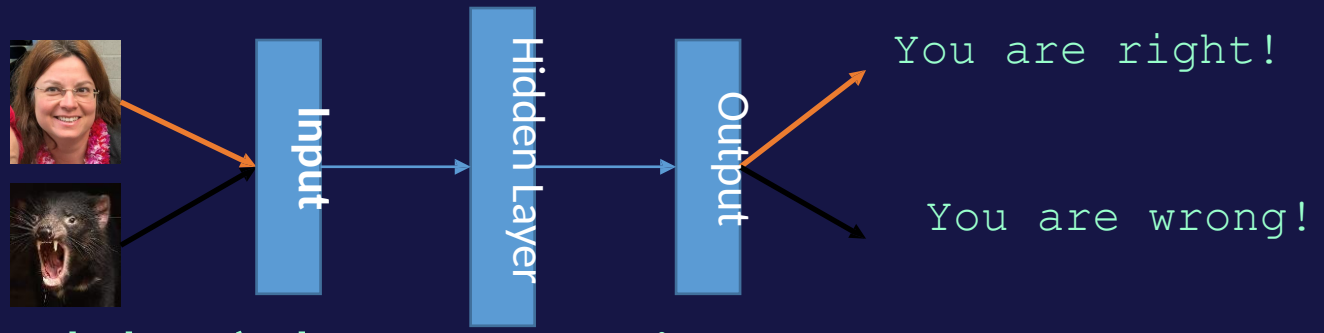
- 'black-box' deep NN:

What is bias in AI?

- Explicit, rule-based AI:

```
IF sees(system, me)
THEN output('You are right!')
IF sees(system, my(archenemy))
THEN greet('You are wrong!')
```

- 'black-box' shallow NN: train on



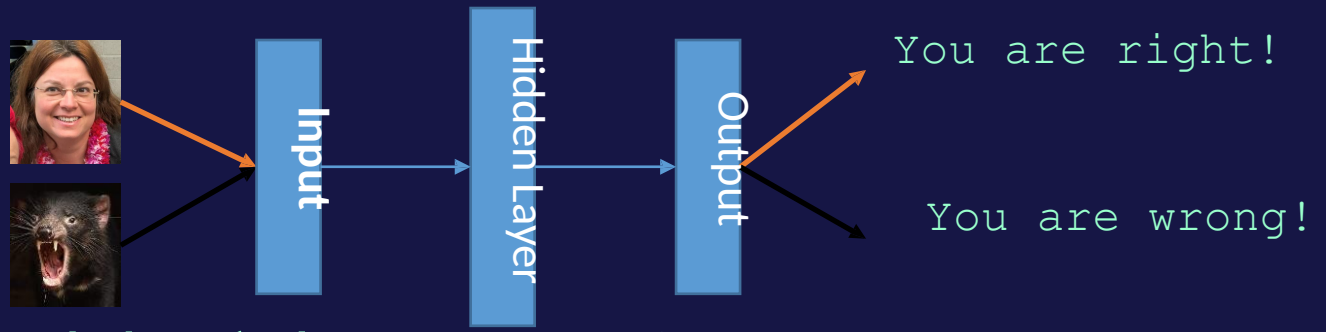
- 'black-box' deep NN: train on

What is bias in AI?

- Explicit, rule-based AI:

```
IF sees(system, me)
THEN output('You are right!')
IF sees(system, my(archenemy))
THEN greet('You are wrong!')
```

- 'black-box' shallow NN: train on



- 'black-box' deep NN: train on