Issues in Machine Learning

Zhiyao Duan
Associate Professor of ECE and CS
University of Rochester

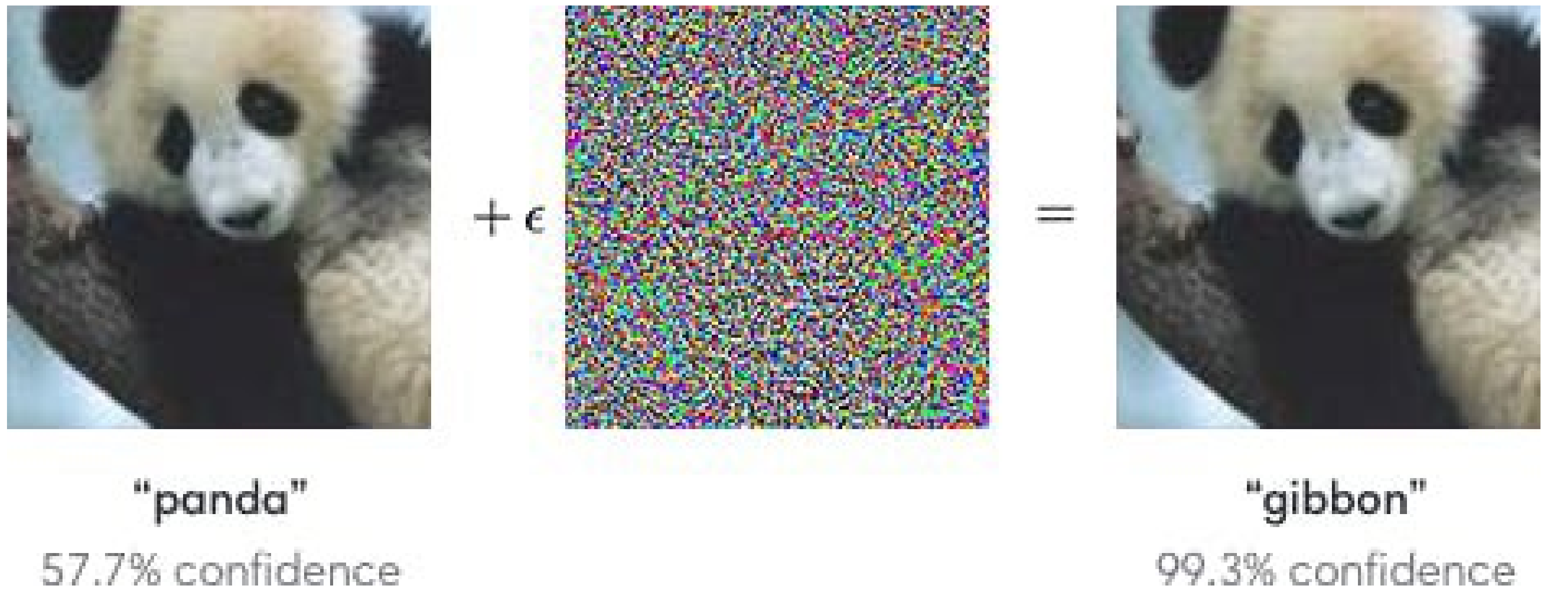
Some figures are copied from the following book

- **LWLS** - Andreas Lindholm, Niklas Wahlström, Fredrik Lindsten, Thomas B. Schön, *Machine Learning: A First Course for Engineers and Scientists*, Cambridge University Press, 2022.

Many Issues

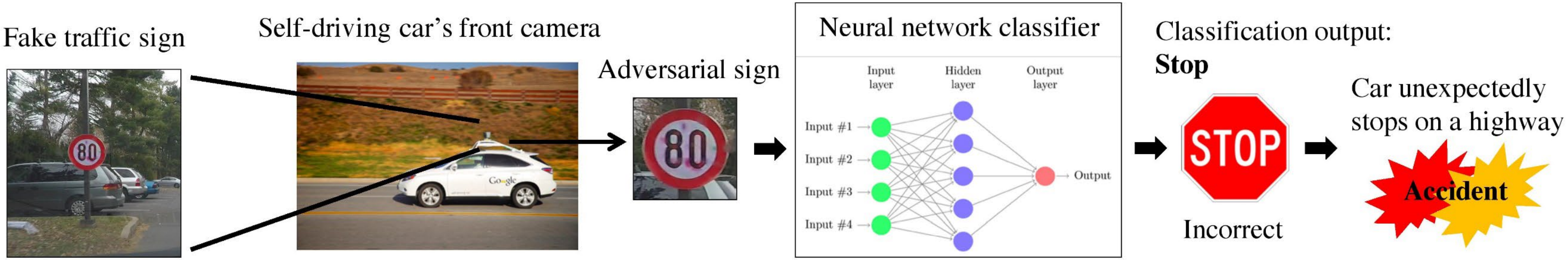
- Robustness
- Explainability
- Accountability
- Fairness
- Bias
- …

Robustness – Adversarial Attacks



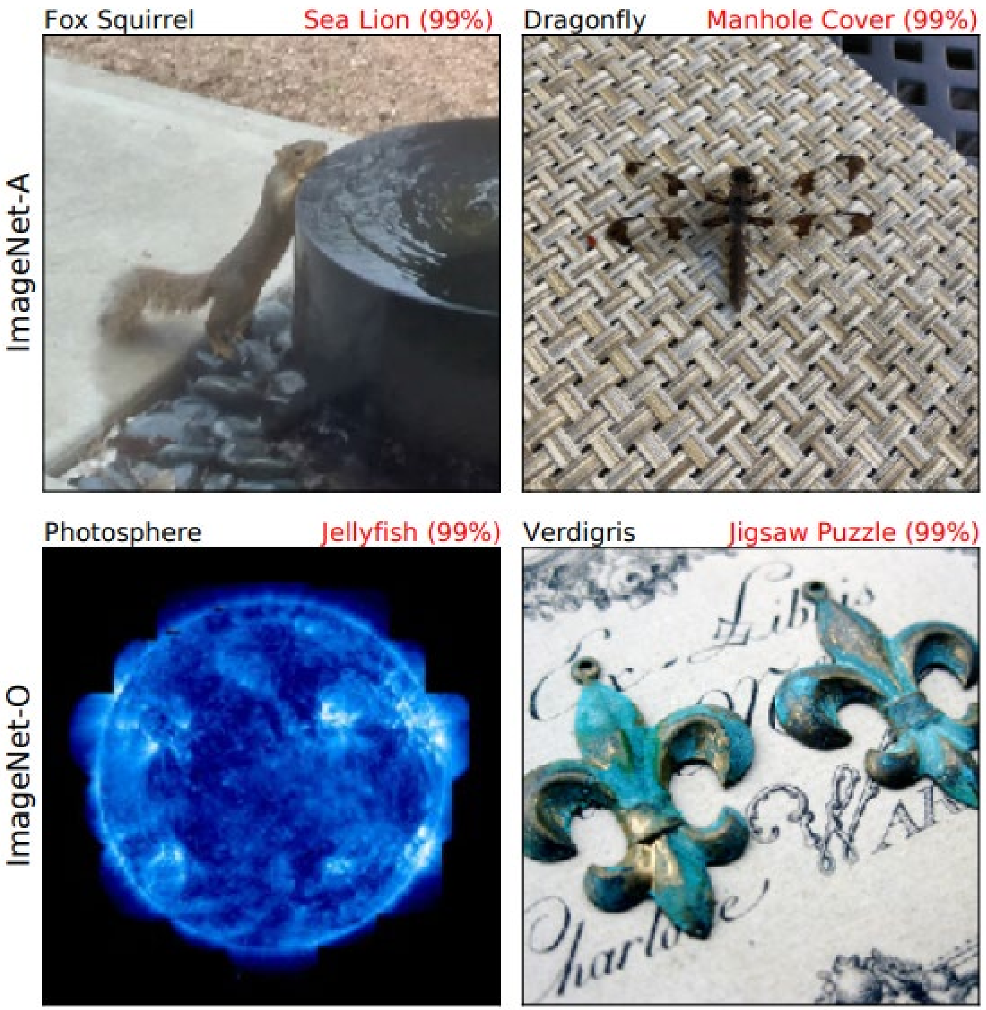
Goodfellow, Shlens, & Szegedy, Explaining and Harnessing Adversarial Examples, 2017.

Robustness – Adversarial Attacks



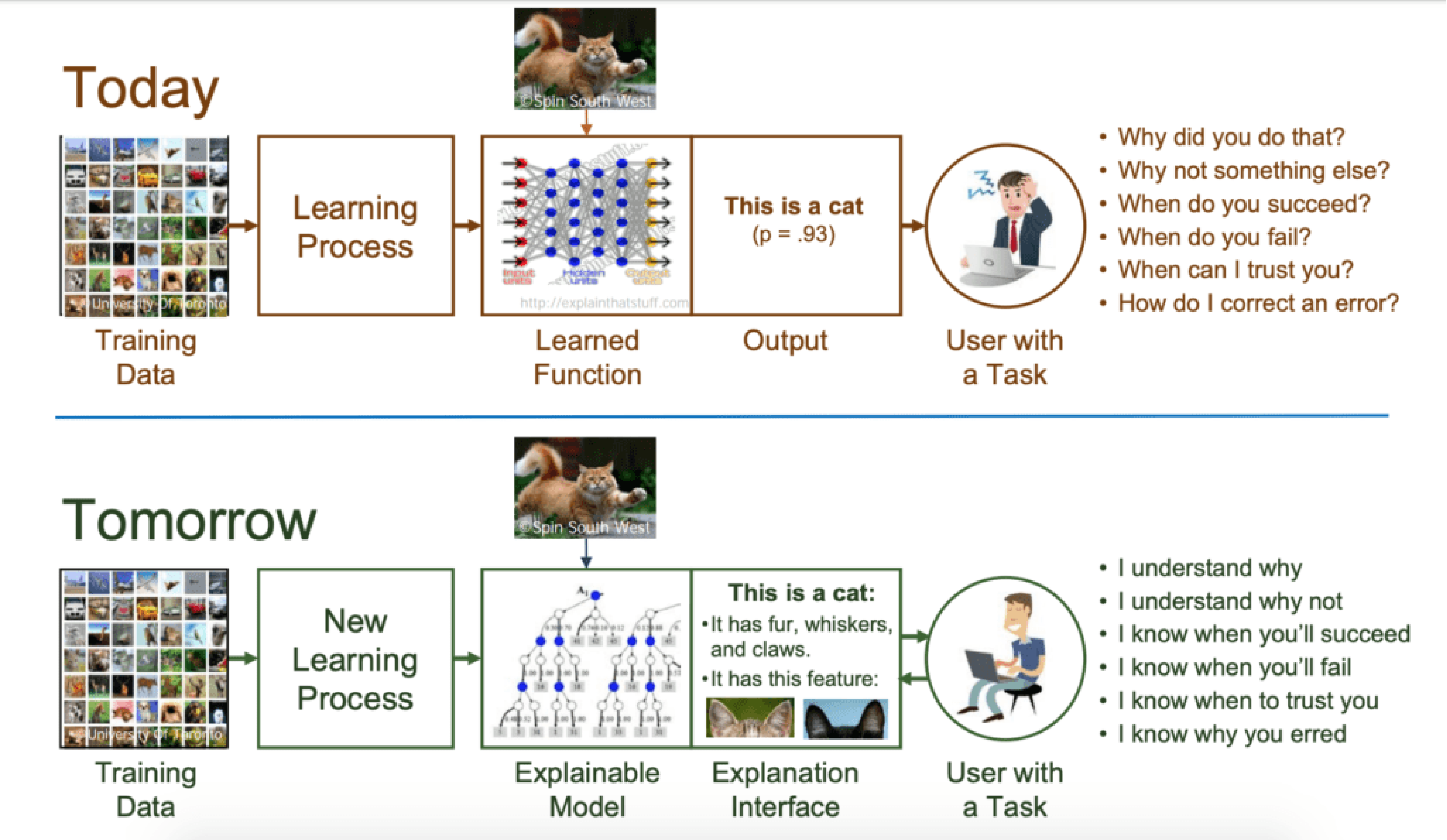
https://adversarial-learning.princeton.edu/darts/

Robustness – Natural Adversarial Examples



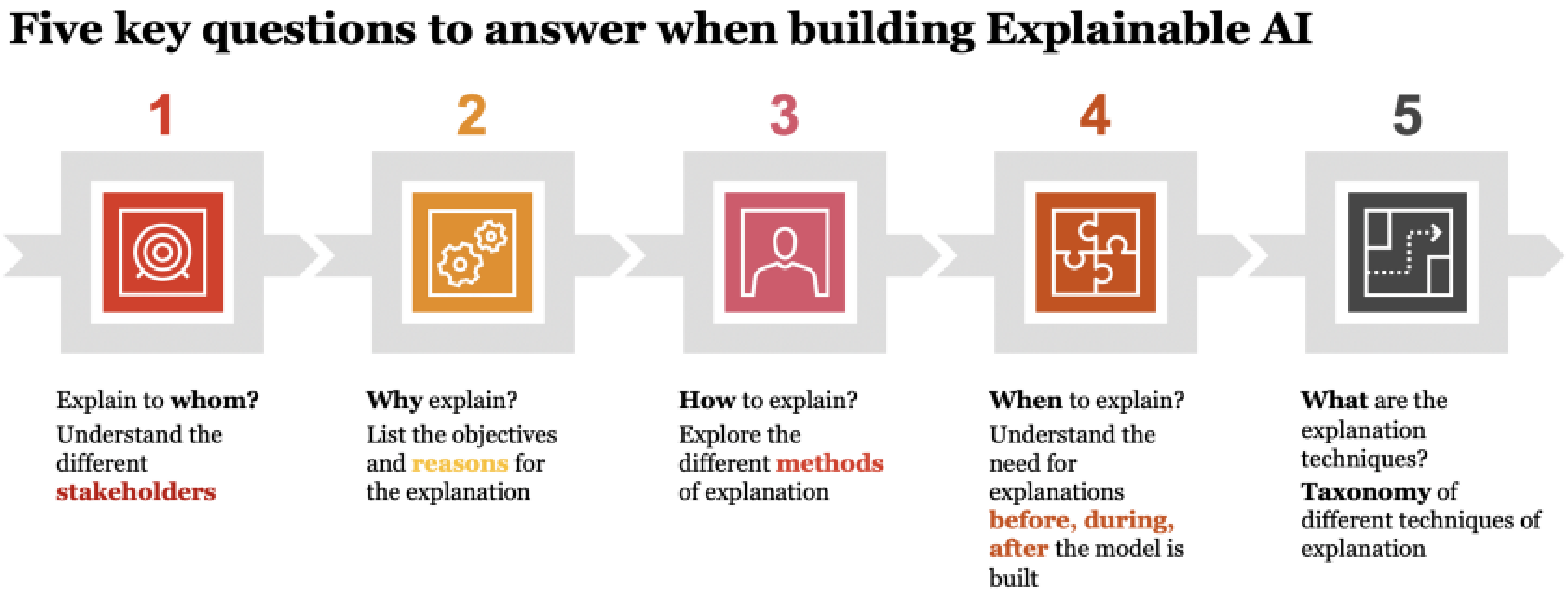
https://arxiv.org/pdf/1907.07174.pdf

Explainability



https://lawtomated.com/explainable-ai-all-you-need-to-know-the-what-how-why-of-explainable-ai/

Explainability



https://swisscognitive.ch/2021/08/23/explainable-ai/

- Avoid misleading claims
  - Exaggeration for commercial gain, cherry-picking results, etc.
- Explain models in understandable ways
  - Avoid jargons

Accountability and Transparency

- External accountability: users/regulators can hold an organization responsible for harmful ML
- Internal accountability: developers/researchers can "debug" a harmful ML system
- Transparency: decisions around fair ML can be understood by stakeholders

Raji et al., *Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing*. https://doi.org/10.1145/3351095.3372873

Fairness

- How to define fairness?
  - Misclassification error rate
    - Non-Swedes: 1/3
    - Swedes: 1/3
  - False Negative Rate (FNR) = FN/(TP+FN)
    - Non-Swedes: 1/2
    - Swedes: 1/9
  - False Positive Rate (FPR) = FP/(TN+FP)
    - Non-Swedes: 1/4
    - Swedes: 7/15

- Definition depends on the application scenario
  - Medical diagnosis vs. criminal sentencing

Table 12.1: Proportion of people shown and/or interested in a course for an imagined machine learning algorithm. The top table is for non-Swedes (in this case we can think of them as citizens of another country, but who are eligible to study in Sweden); the bottom table is for Swedes.

Non-Swedes	Not Interested ($y=-1$)	Interested ($y=1$)
Not recommended course ($\widehat{y}(\mathbf{x})=-1$)	TN = 300	FN = 100
Recommended course ($\widehat{y}(\mathbf{x})=1$)	FP = 100	TP = 100

Swedes	Not Interested ($y=-1$)	Interested ($y=1$)
Not recommended course ($\widehat{y}(\mathbf{x})=-1$)	TN = 400	FN = 50
Recommended course ($\widehat{y}(\mathbf{x})=1$)	FP = 350	TP = 400

Fairness

- How to define fairness?
  - False Positive Rate (FPR) = FP/(TN+FP): Not fair
    - Black: 805/(990+805) = 44.8%
    - White: 349/(1139+349)=23.4%
  - True Positive Rate (TPR) = TP/(FN+TP) = Recall: Not fair
    - Black: 1369/(532+1369)=72.0%
    - White: 505/(461+505)=52.2%
  - Precision = TP/(TP+FP): OKay
    - Black: 1369/(1369+805)=63.0%
    - White: 505/(505+349)=59.1%

Table 12.2: Confusion matrix for the Pro-Publica study of the Compas algorithm. For details see Larson et al. (2016).

Black defendants	Didn't reoffend ($y = -1$)	Reoffended ($y = 1$)
Lower risk ($\widehat{y}(\mathbf{x}) = -1$)	TN = 990	FN = 532
Higher risk ($\widehat{y}(\mathbf{x}) = 1$)	FP = 805	TP = 1 369
White defendants	Didn't reoffend ($y = -1$)	Reoffended ($y = 1$)
Lower risk ($\widehat{y}(\mathbf{x}) = -1$)	TN = 1 139	FN = 461
Higher risk ($\widehat{y}(\mathbf{x}) = 1$)	FP = 349	TP = 505

Fairness

- Theorem: if we cannot perfectly classify the data and the base rate (positive/negative) of the outcome differs between the two groups, then it is impossible to achieve simultaneous equality (i.e., fairness between groups) in precision, true positive rate (recall), and false positive rate!
  - If we achieve equality in two of them, then the third one must not equal!

- We should be aware of these limitations and explain them to users

- Kleinberg, Jon, et al. 2018 "Algorithmic fairness."
- Chouldechova, Alexandra, and Aaron Roth. 2018 "The Frontiers of Fairness in Machine Learning."

Bias

- Word2vec: learned embeddings for words

$$\mathrm{Water} - \mathrm{Liquid} + \mathrm{Gas} = \mathrm{Steam},$$

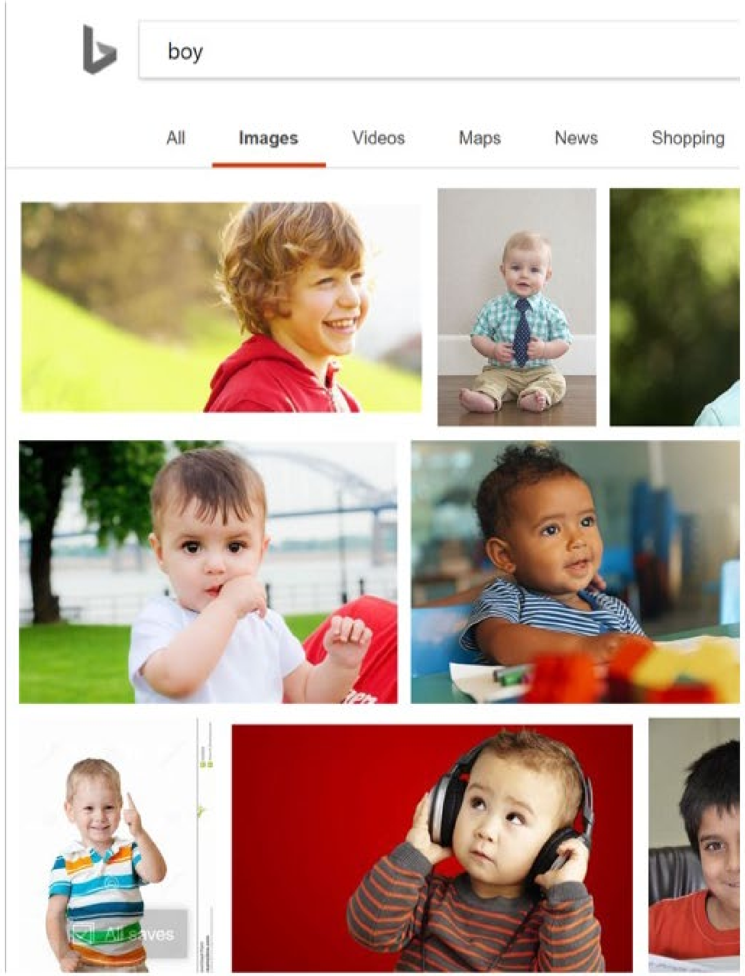
$$\mathrm{Intelligent} - \mathrm{David} + \mathrm{Susan} = \mathrm{Resourceful}$$

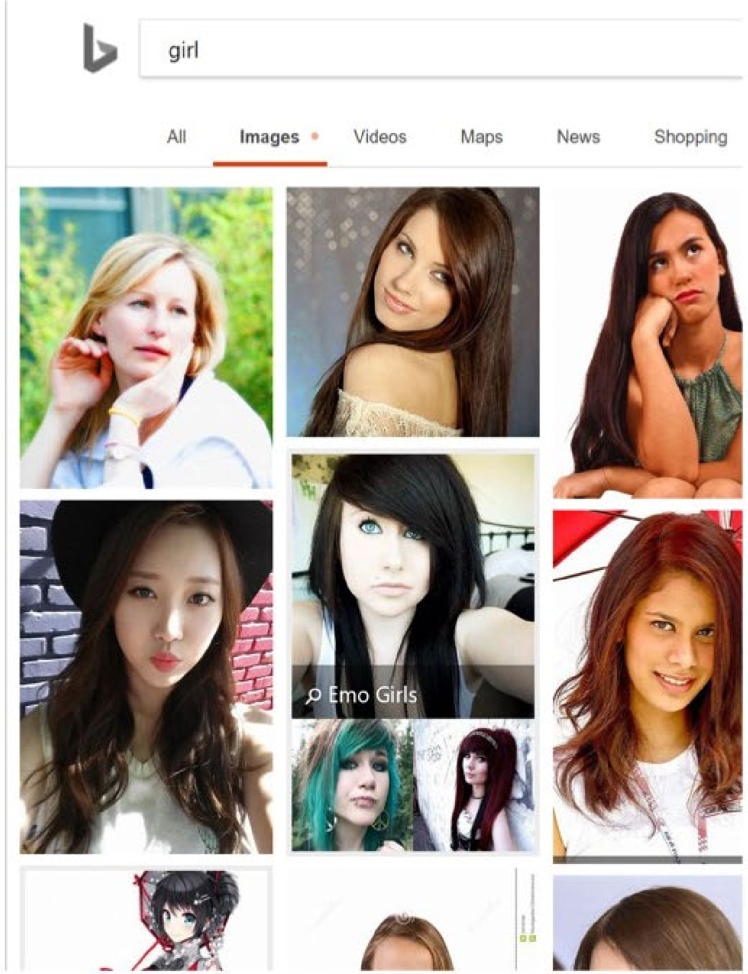
$$\mathrm{Brainy} - \mathrm{David} + \mathrm{Susan} = \mathrm{Prissy}$$

$$\mathrm{Smart} - \mathrm{David} + \mathrm{Susan} = \mathrm{Sexy}$$

Bolukbasi, Chang, Zou, Saligrama, & Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," NIPS 2016.

Bias


boy
All
Images
Videos
Maps
News
Shopping


girl
All
Images
Videos
Maps
News
Shopping
Emo Girls

Other Issues

- Privacy
  - We are losing privacy as AI models advance
  - We are more vulnerable to data misuse
  - Information remains in models even if data is deleted
- Copyright
  - Generated content can be very similar to copyrighted content
- Sustainability
  - Sustainability of machine learning: as models become larger, energy consumption (and carbon footprint) increases
  - Machine learning for sustainability
- Misuse and abuse
  - Open-source pros and cons
  - Regulation on AI development?