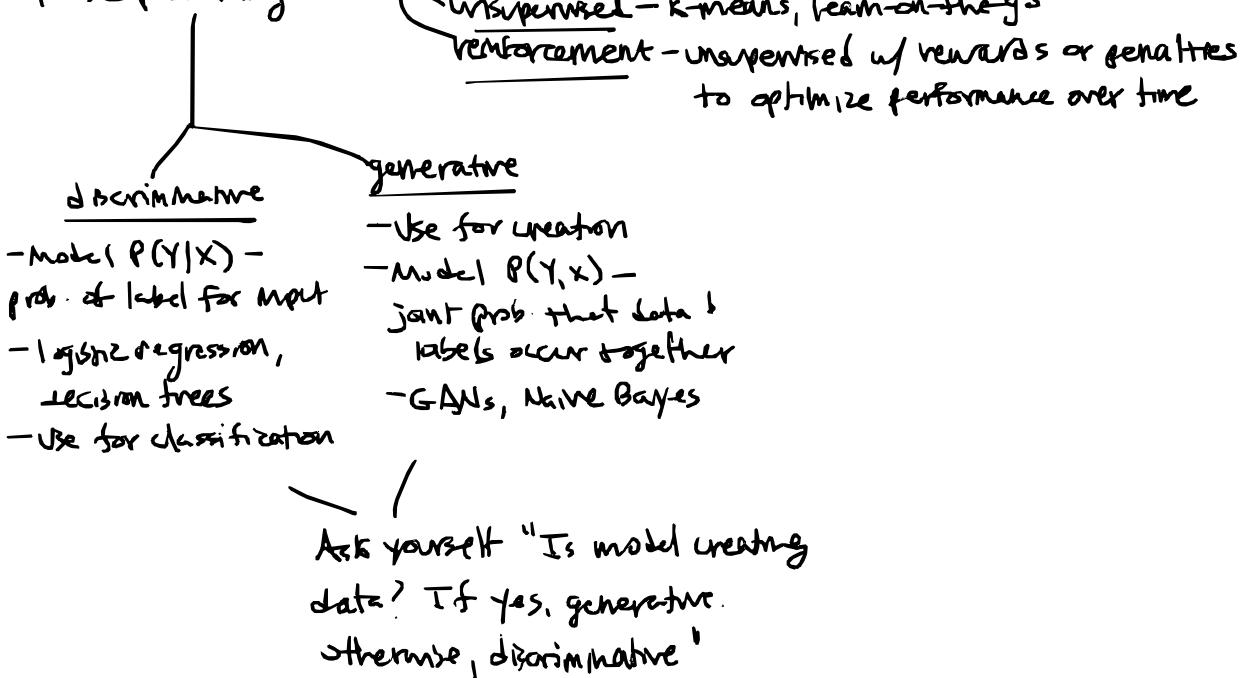


1. Symbolic / Rule-based AI ("Expert Systems")
 - GOFAI (Good Old-Fashioned AI)
 - Uses explicit symbols, rules, & logic to perform reasoning
2. Modern AI are built off neural networks.
 - Model AI off human brain.
 - Nodes in neural nets are associated w/ dynamically adaptable weights, which change w/ inputs to system
3. Shallow learning vs. Deep Learning
 - few layers
 - many layers
 - in neural networks
4. Deep Learning
 - Supervised - learn using training data; then extrapolate
 - unsupervised - k-means, learn-on-the-go
 - reinforcement - unperfected w/ rewards or penalties to optimize performance over time



5. Features
 - (input/activation layer) → hidden layers → output layer

W/ deep learning, we have hundreds/thousands of hidden layers, w/ dynamically adjusting weights.

6. Word embeddings - representation of text where words or phrases from vocab. are mapped to vectors of real #s

↳ words w/ similar meanings are closer together.
Those w/ diff meanings are further apart.

7. Transformer Models - processes word embeddings

↳ use attention mechanisms to understand rels b/w words in a sentence

8. Gen. AI systems are actually predictive systems.

↳ "stochastic parrots" - They repeat things w/o true comprehension.

9. Limits of algorithmic systems

- Related to 3 key attributes

↳ data-driven
↳ base on statistical models
↳ use dynamic rule development

- Data-Driven

* Data is not reality. It is a representation/model of reality to focus on the aspects of reality closest to the problem we're solving.

* Data requires quantification.

* Any representation of data is a choice - and can leave things out.

- Stat. Models

* Fundamentally correlational

* pattern recognition engines

* No understanding of causal relationships

10. AI Accuracy

- Capability: Meets AI's defined objective fn. → if a classifier is producing wrong classifications

- Alignment: Meets intended human objective

↳ includes things like slope for LMs → Don't use customer service bot for medical diagnosis.

- Robustness: May be accurate at launch --- but does it stay accurate?

toxicity

11. Oversfitting - Model has basically memorized the training data instead of learning underlying patterns.
↳ for anything outside training data, it will perform horribly.
TLDR: oversetting runs generalization.

12. Distinctive Generative AI Accuracy Issues

- Unpredictable applications
- Hallucinations
 - ↳ Not related to existing examples in dataset
- Concentration risk of train. models
 - ↳ You end up getting just a few companies owning & running actual train. models, just like w/ hyperscalers
- Need to have room for error? randomness b/c generative AI is creative - b/c it's creating data
- PEAKAC (problem exists b/c keyboard and chair)
 - ↳ People = source of error too but still an integral part of sys!

13. Transparency — Awareness: Is an AI system being used?

- ↳ Disclosure: what are the features/attributes?
what are the model weights, tuning params,
gradients, etc.?

Interpretation or Explanation

why did the system produce
the result that it did?

14. Improve transparency thru
- ↳ algorithmic audits
 - ↳ model/system cards — vendor publishes notes
on model + how it works + was built
5th, 2nd, 3rd party

15. Interpretable - linear models subject to direct understanding
 - ↳ Deep learning / LMs = not interpretable b/c of many hidden layers

16. Explainable - finding interpretable models consistent w/ the output

17. Limit of Interpretability / Explainability

- Technical Illiteracy
- Not a description of the actual "black box"
 - ↳ just an approximation

18. You can ask a LLM to explain itself... but reliability is not guaranteed.

- Buyer beware - LLM fundamentally non-deterministic, so it can lie and change response to prompt for explanation every time

19. Common Means of Explainability — using technical tools to describe model's features / workings

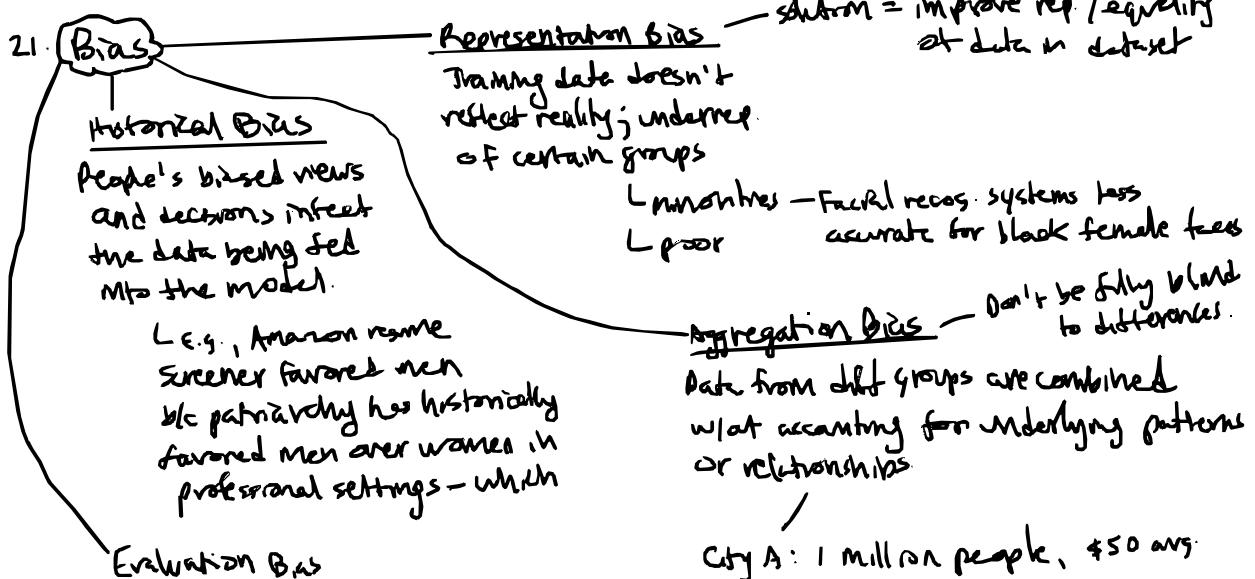
/

Compare to simpler model w/ similar objective fnn.

Companion model to predict & explain complex math model's outputs

20. Label Choice Bias — Gap b/w ideal (what we want) vs actual (what it actually does)

- ↳ E.g., AI models prioritizing minorities or reducing cost of care by disproportionately most vulnerable



$$AB \rightarrow \text{If you say avg. income} = \frac{50 + 30}{2} = 40$$

$$\text{B/c actual avg} = \frac{50(1e6) + 30(1e0)}{1e6 + 1e0}$$

Avg has to take # of people in each city into account.

Mechanisms used to evaluate AI themselves have to introduce bias.

- ↳ All facial recognition systems trained off / optimizing from common eval database t. — which was massively biased

22. Bias doesn't just come from data - also from human choices in creating & deploying the model.
23. Deepfakes - identity theft, where bad actor uses AI system to pretend to be someone else in order to fool/exploit people
24. Synthetic data - using AI to create data that is then used to train other AI
25. In the US, generative AI outputs are not protected by copyright.
 - Copyright only protects human authorship.
 - Beijing court disagrees - beginning of a different task in China
 - Tension b/w protecting copyright/IP vs. promoting AI innovation
26. Licensed/synthetic/public-domain data → Don't have to worry about copyright restrictions
27. Privacy Lifecycle
 - Collection/Creation → what data to gather → how
 - Prioritize data minimization (only gather what's needed)
 - Aggregation & Analysis - Data combined + processed
 - Use - Data used to power apps/services
 - Storage - Is it safely stored? Encryption at rest / backups / access control
 - Distribution - Who is it shared with?
28. Metaprogramming - "Here are the rules! patterns to follow when answering questions like this"
 - └ "you are ..." - giving a person a guideline for how to behave