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 ntwrk are associated w/ dynamically adaptable weights, which change w/ inputs to system

3. Shallow learning vs. Deep Learning
   - shallow learning: few layers
   - deep learning: many layers
   - in neural networks

4. Deep Learning
   - supervised – learn using training data & then extrapolate
   - unsupervised – k-means, learn-on-the-go
   - reinforcement – unsupervised w/ rewards or penalties to optimize performance over time

   discriminative
   - model $P(Y|X)$ – prob. of label for input
   - logistic regression, decision trees
   - use for classification

   generative
   - use for creation
   - model $P(Y, X)$ – joint prob that data & labels occur together
   - GANs, Naive Bayes

   Ask yourself "Is model creating data? If yes, generative. otherwise, discriminative"

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
     - based on statistical models
     - use dynamic rule development
   - Data-Driven
     - Data is not reality. It is a representation / model of reality to have in 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 fxn — If a classifier is producing wrong classifications
    - Alignment: Meets intended human objective
      - includes things like scope for LLMs — Don't use customer service bot for medical diagnosis.
      - toxicity
    - Robustness: May be accurate at launch --- but does it stay accurate?

11. Overfitting – Model has basically memorized the training data instead of learning underlying patterns.
↳ for anything outside training data, it will perform horribly.

TLDR: overfitting ruins generalization.

12. Distinctive Generative AI Accuracy Issues
- Unpredictable applications
- Hallucinations
  - Not tied to existing examples in dataset
- Concentration risk of fndn. models
  - You end up getting just a few companies owning & running actual fndn. models, just like w/ hyperscalers.
- Need to have room for error & randomness b/c generative AI IRL is to be creative – b/c it's creating data
- PEBKAC (Problem exists btw keyboard and chair)
  - People = source of error too but still an integral part of sysl

13. Transparency
- Awareness: Is an AI system being used?
- Disclosure: What are the features/attributes? What are the model weights, tuning params, guardrails, etc.?
- Interpretation or Explanation: Why did the system produce the result that it did?

14. Improve transparency thru
- algorithmic audits – 1st, 2nd, 3rd party
- model/system cards – vendor publishes notes on model & how it works + was built

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

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

17. Limits 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 give and change response to prompt for explanation every time

19. Common Means of Explainability
  - using technical tools to describe model's features & workings
    - Companion model to predict & explain complex main model's outputs
  - Compare to simpler model w/ similar objective fxn

20. Label Choice Bias – Gap btw ideal (what we want) vs actual (what it actually does)
  - e.g., AI models directing ambulances or reducing cost of care by disqualifying most vulnerable

21. Bias
  - Representation Bias
    - Training data doesn't reflect reality; underrep of certain groups
      - minorities – Facial recog systems less accurate for black female faces
      - poor
    - solution = improve rep / equality of data in dataset
  - Historical Bias
    - People's biased views and decisions infect the data being fed into the model
      - e.g., Amazon resume screener favored men b/c patriarchy has historically favored men over women in professional settings – which
  - Aggregation Bias – Don't be fully blind to differences
    - Data from diff groups are combined w/o accounting for underlying patterns or relationships
      - City A: 1 million people, $50 avg
        City B: 100 people, $60 avg
      - AB → If you say avg. income = $\frac{50+30}{2} = 40$
      - B/c actual avg = $\frac{50(1e6) + 30(100)}{1e6 + 100}$
      - Avg has to take # of people in each city into account
  - Evaluation Bias
    - Mechanisms used to evaluate AI themselves have & introduce bias.
      - All facial recognition systems trained off / optimizing from common eval dataset. — 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 track in China
   - Tension btw 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. Metaprompting – "Here are the rules & patterns to follow when answering questions like this"
   - "You are ..." – giving a persona as guideline for how to behave