Definition, History, and Development of AI

A few key summary points

- AI is a foundational technology that is advancing other scientific fields and has the potential to transform how society operates like electricity and the internet.
- Even the most advanced AI has many failure modes that are unpredictable, not widely appreciated, not easily fixed, and not explainable, and that can lead to unintended consequences.
- The approach to AI with the most positive impact on employment calls for the AI-enabled augmentation of human capabilities rather than the replacement of humans by machines.
- Failure of policy makers to act will guarantee bad outcomes will be realized.
- Be very wary of AI “snake oil” and ask probing questions. Technology advocates have a way of sweeping hard questions under the rug.

- Multiple, varied, and often inconsistent with each other.
  - AI is “the science and engineering of making intelligent machines.” What is intelligence?
    - Ability to answer questions at human levels of competence
    - Ability to solve problems at human levels of competence
    - Ability to achieve goals at human levels of competence
    - Notice what is omitted: asking good questions, formulating useful problems, defining meaningful goals. Also, empathy, compassion, justice, fairness
  - Artificial general intelligence
    - AI that can (learn to) accomplish any intellectual task that human beings can perform
    - AI that can surpass human performance on most economically valuable tasks.
  - Artificial narrow intelligence
    - Problem specific rather than general
      - Doing arithmetic and math calculations, chess, speech transcription were all once “hard”

How we got here

- Short history
  - Turing test (1950)
  - Perceptrons
  - Reasoning as search through possibilities
  - Natural language understanding and translation
  - Expert systems
- Why now? Convergence of several trends
  - Sufficiently powerful hardware (Processing units, memory)
  - Large-scale availability of data (all of the Internet to mine)
  - New approaches
    - Data driven (which finds patterns) rather than algorithmically driven (which assumes patterns)
    - Neural networks (Act II)
    - Probabilistic reasoning

State of today’s art

- Some subfields of AI:
  - **Machine learning**, enabling computers to perform tasks without explicit instructions, often by generalizing from patterns in data.
  - **Computer vision**, enabling machines to recognize and understand visual information from the world, converting it into digital data and making decisions based on it
  - **Natural language processing**, equipping machines with capabilities to understand, interpret, and produce spoken words and written texts
- Machine learning:
  - Supervised learning: using labeled input data (e.g., images labeled as “cats” or “dogs,” house prices with various features), making classifications or quantitative estimates
  - Unsupervised learning: using unlabeled data, identify patterns and structure in the data without supervision, e.g., clustering similar data together based on their features.
  - Reinforcement learning: learning optimal behavior based on training through rewards and penalties as trial-and-error interactions with the environment occur

Some applications today

- Health Care
  - Medical diagnostics (automated triage for strokes)
  - Drug discovery (identifying promising drug compounds)
  - Robotic assistants. (in-hospital deliveries, assisting physical therapists)
- Agriculture
  - Production optimization (sorting salmon)
  - Crop management (optimal use of pesticides)
- Logistics and Transportation
  - Resource allocation (prediction of shipping times)
  - Predictive maintenance
- Law
  - Legal transcription.
  - Legal review.

In recent news: ChatGPT and other LLM-based chatbots

- Write a customized resume and cover letter
- Create original jokes
- Explain complex topics
- Solve math problems step-by-step
- Give relationship advice
- Write music in almost any genre
- Write, debug, and explain code
- Brainstorm and generate ideas
- Translate text
- Create, edit, and modify media files
- Make recommendations
- Play games
- Get cooking help
- Improve personal health
- Assist with travel plans
- Prepare for a job interview
- Write essays on almost any topic
- Help with writing
- Summarize documents

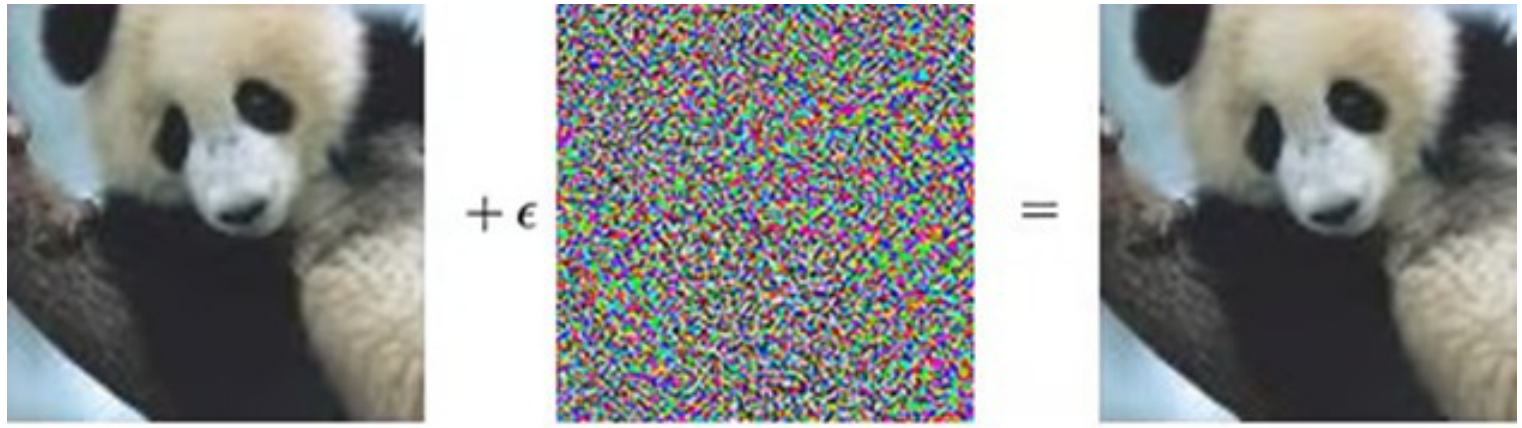
Two key terms

- Large language models (LLMs)
  - trained on very large volumes of written text to recognize, summarize, and generate new text, based on a statistical analysis that makes predictions about what other words are likely to be found immediately after the occurrence of certain words.
- Chatbot
  - Computer system that interacts with humans conversationally in natural language.

Flies in the ointment

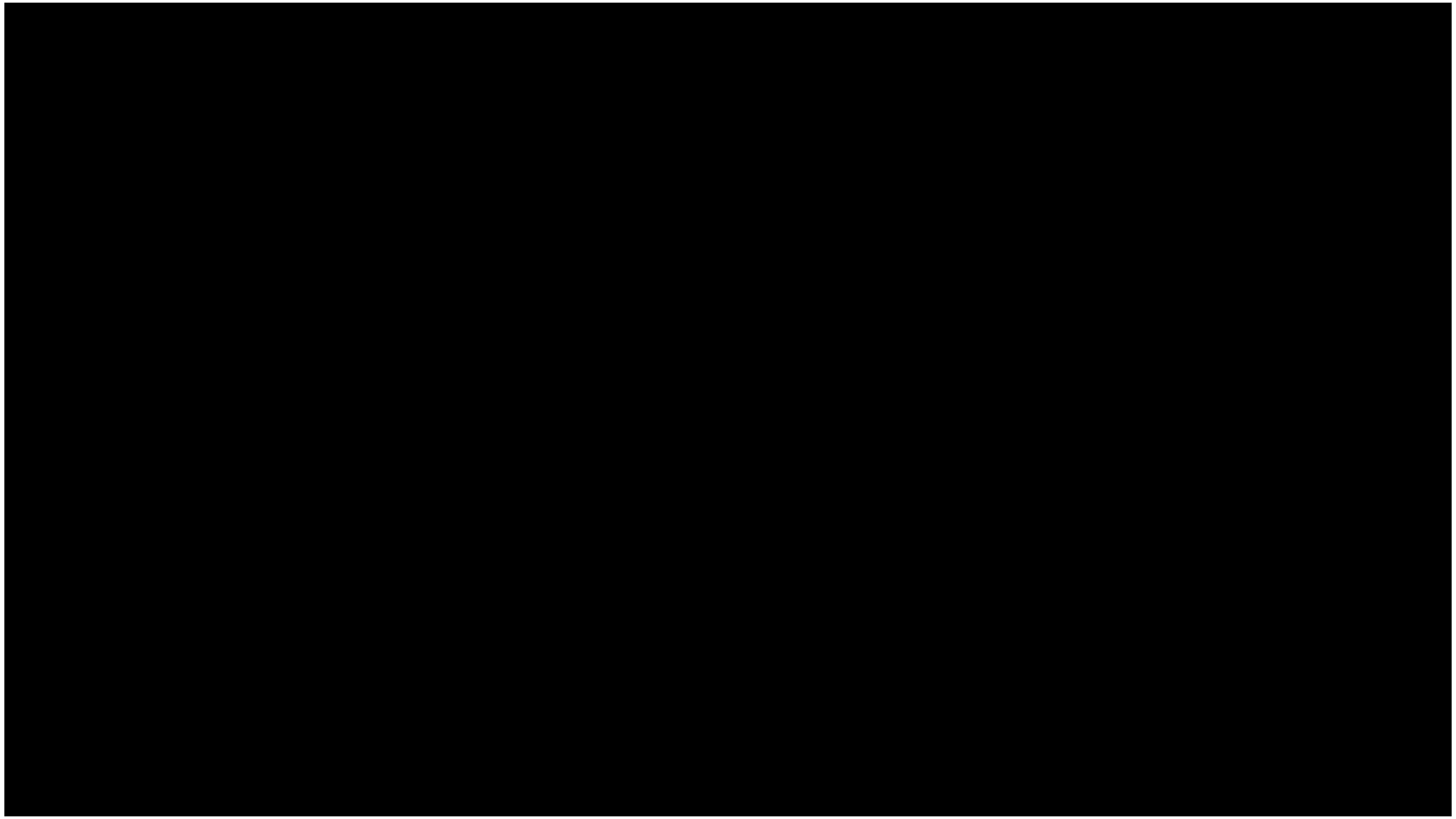
- Explainability:
  - AI generally incapable of explaining the basis on which it arrives at any particular conclusion.
  - Explanations are not always relevant but in certain cases may be critical.
  - Safeguards against bias of various kinds
- Bias and fairness.
  - Bias is a property of the data that is commonly regarded as societally undesirable.
  - If bias is reflected in training data, output based on training data will reflect such biases.
    - History: male/female, doctor/nurse
    - Facial recognition: training facial recognition on one ethnic group may reduce accuracy in identifying people from other ethnic groups.
  - Addressing bias:
    - Changing the training data
    - Human coding of guardrails into the output

- Spoofing (https://arxiv.org/pdf/1412.6572.pdf)



- Overtrust: excessive trust placed in machine output.
- Hallucinations: generation of plausible but incorrect results
- Out-of-distribution (OOD) inputs: inputs too different from training data

Deepfakes



On the future of work

- Individuals whose jobs entail routine white-collar work may be more affected than those whose jobs require physical labor
  - Some painful shifts in the short term are inevitable.
- AI is helping some workers to increase productivity and job satisfaction.
- Other workers are already losing their jobs as AI—in some cases despite underperforming humans—demonstrates adequate competence for business operations.
- Training displaced workers to be more competitive in an AI-enabled economy does not solve the problem if new jobs are not available.
  - The nature and extent of new jobs are not clear at this point, although historically the introduction of new technologies has not resulted in a long-term net loss of jobs.

Many codes of responsible AI use

- The emphasis is on creating AI that accounts for human welfare, dignity, equity and social good.
  - Beneficence and non-maleficence - AI systems should provide significant benefit to humanity while ensuring they do not cause physical or social harm.
  - Justice, fairness and inclusion - AI must not create or exacerbate unfair bias and should promote equitable access and inclusion for all regardless of identity or social status.
  - Accountability and oversight - Appropriate accountability and governance structures must exist to ensure AI systems act ethically. This includes transparency, explainability, human control, and evaluation mechanisms.
  - Respect for human rights and dignity - AI oversight must respect fundamental human rights and human dignity. This includes privacy, autonomy, freedom of choice, as well as solidarity and human flourishing.
  - Reliability and trustworthiness - Rigorous engineering and validation of AI systems should engender confidence they will behave as intended in deployment contexts.
- Easy to say, hard to do.

Opinion

- The cynic’s view of AI: AI is something that
  - Sometimes or even mostly works but **not** all of the time
  - When it won’t work,
    - is unpredictable, and
    - perhaps unnoticeable
  - When it doesn’t work,
    - You don’t know why it didn’t work.
    - You don’t know how to fix it except by fiddling or adding a brute-force ex post facto rule.
- Your job as legislators is not to accept at face value what technology vendors and advocates tell you.
  - The fundamental question you must ask: how and to what extent, if any, have you (the vendor or advocate) subjected your system to dedicated, no-holds-barred adversarial testing intended to break it?
  - One bit of advice: crawl before walk, walk before run. Do low-risk applications first.