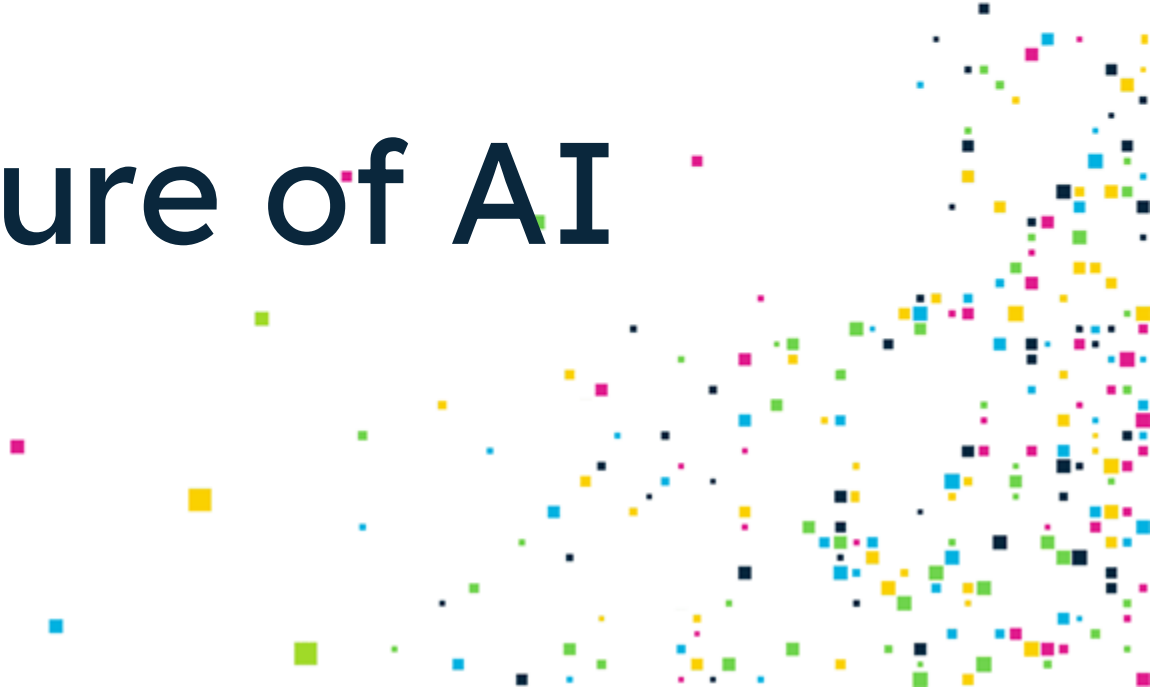
Future of AI



can code communities

Review Take Home

Find one example of AI ethics being used within real world application OR one example of where AI ethics should be introduced and how.

Review Questions

- What is the difference between implicit and explicit bias?
- What is one of the multiple AI safety concerns that we face today?
- What is one of the techniques used in Explainable AI?

Future of AI

- In an era marked by the rapid advancement of AI, it's imperative to not only grasp the current state of AI technology but also anticipate its future trajectory.
  - Businesses
  - Policymakers
  - Individuals
- By exploring the latest trends, challenges, and opportunities in AI, participants will gain the knowledge and insights necessary to navigate this dynamic field with confidence.

Consumer Experience

- AI is being seamlessly integrated into our daily lives, shaping our expectations along the way.
- AI is everywhere, with 77% of devices in use featuring some form of AI.
    - Personalized recommendations on streaming platforms to voice assistants guiding us through tasks
- Currently, 50% of consumers remain optimistic about AI, recognizing its potential to enhance daily tasks.
    - This number will only grow in the upcoming years, which will lead to more efficient AI adoption.
- In light of these insights, it's evident that user experience plays a crucial role in shaping realistic expectations of AI.
    - By understanding how users interact with AI in their everyday lives, developers can design solutions that resonate with users and deliver tangible value.

Business Adaptations

In parallel, businesses are increasingly embracing AI, with 35% having already adopted it. Notably, 9 out of 10 organizations view AI as essential for gaining a competitive advantage, highlighting its transformative potential.

- Notion
  - It auto-categorizes notes, suggests project timelines, and even drafts summaries from meeting notes.
  - This AI integration makes project management and personal organization seamless and intuitive, saving time and boosting productivity.

Business Adaptations

- Adobe
  - The latest version of Adobe Creative Cloud includes new generative AI features that put even more innovation in the hands of creators of all types.
  - These features include generative fill in photoshop, generative b-roll in premiere pro, and many more.
- Shopify
  - The e-commerce company combined all its mighty powers with the latest advancements in AI to provide more personalized support for various tasks—from copy generation to store building, marketing, and customer support.
  - Shopify Magic was introduced: a suite of free AI-enabled features integrated across Shopify's products and workflows to make it easier for users to start, run, and grow their businesses.

Standalone AI

- While there are many companies integrating AI into their own products or services, there are also companies that are pushing the bounds of standalone AI products.
- Most of the current examples show a lot of potential for changing the job market and our lifestyles, but don’t deliver the level of performance necessary for those changes.

Devin AI

- This AI software engineer product
  - Similar to an autocoder
  - Has control of common developer tools including the shell, code editor, and browser within a sandboxed compute environment.
- After the initial announcement, a lot of developers were worried about this product taking their jobs from them.
  - Fortunately we are still a ways out from that happening.
- Devin has a limited task range, takes a noticeable amount of time and tokens to complete tasks, and still requires human oversight.

Figure O1

- Advertised as the world's first commercially-viable autonomous humanoid robot, which is powered by the GPT family of language models.
- The result is a robot that can "see", respond to external stimuli, and converse fluently with humans in its vicinity.
- The biggest drawback of this product is reaction latency.
    - Since the GPT models are huge language models, the robot captures audio and/or image and sends the input data over the internet for computations to be done on a cloud service provider.
    - This creates a lot of latency between the prompting action and the reaction of the robot, which is far from ideal.
    - Think about the tasks you would want an AI powered robot to perform, and how issues could arise due to the latency.

AI Pin and Rabbit R1

- Both of these products have similar utilization and drawbacks.
- They are marketed as AI assistants that can be brought with users wherever they are.
- The idea of having AI be able to assist users throughout the day outside of the normal computer environment is great, but there are plenty of factors that make these products currently undesirable.
    - Just like figure 01, these devices rely on cloud services to compute responses, which adds latency to the experience.
    - The reliance on cloud services also requires these products to be connected to the internet, whether Wi-Fi or cellular data.

AI Pin and Rabbit R1

- These requirements make the device description very similar to a smartphone, but with less capabilities.
  - Most smartphones have AI assistants integrated into their system already (Siri and Google Assistant), and the ones who don’t can use the ChatGPT mobile app to perform similar tasks.
  - 92% of Americans own smartphones and have all their apps all connected within that device, so adding another device with less capability seems useless.

Exercise

Search and find a recent announcement of an AI (standalone or incorporated) product and discuss the potential benefits of the product as well as the limitations that this product might face.

Trends in AI

- Currently there is a gap between what is expected of AI products and their performance, but that will change over time and development.
- There are many facets of AI that are trending and have potential to grow the landscape of its capability.

Multimodal AI

- These multimodal models (GPT4V, Gemini, etc.) represent a significant leap forward in AI development, as they integrate multiple input layers, including text, images, and video, to generate more comprehensive and contextually rich outputs.
- This multidimensional approach enables AI systems to understand and interpret complex scenarios with greater depth and sophistication.
- The incorporation of video into AI models represents a significant advancement in the field of artificial intelligence.
    - From recognizing objects and actions to interpreting gestures and facial expressions, video inputs enhance the capabilities of AI models, opening up new possibilities for applications in fields such as video captioning, content creation, and video analysis.

Smaller Models

- Training a model of comparable size to GPT3 requires the yearly electricity consumption of over 1,000 households
- An average day of ChatGPT queries is equivalent to the yearly consumption of 33,000 households.
    - To reduce the resources required and the environmental impact, there has been a focus on smaller sized models.
        - Faster inference times, and lower costs associated with them due to their nature.
        - Storing local models on common devices like laptops.
            - This helps with accessibility and user experience.

Mixtral

- Mixtral is known as a “mixture of experts” model which is composed of 8 neural networks with 7 billion parameters each.
- This 56 billion total parameter count model has outperformed Llama 2 (which has 70 billion parameters) on most benchmarks with an inference speed that is 6 times faster.
- It also matches or outperforms GPT3.5 (which has 175 billion parameters) on most standard benchmarks.

Managing GPU and Cloud Computing

- With the increasing demand for GPUs and the rise of cloud computing, effective cost management strategies are essential for optimizing resources and maximizing efficiency.
  - GPUs excel in parallel processing tasks, making them indispensable for training deep neural networks and accelerating inference tasks.
  - As AI applications become more sophisticated and data-intensive, the demand for GPUs continues to grow exponentially.
- Cloud computing offers scalability, flexibility, and accessibility, making it an attractive option for AI development.
  - However, cloud computing costs can quickly escalate, particularly for resource-intensive tasks like training large AI models.

Optimizing Model Size and Performance

Optimizing model size and performance is essential for reducing GPU and cloud costs while maintaining or improving AI system efficiency.

- Model Pruning
  - Identifying and removing redundant or unnecessary parameters from AI models to reduce size and improve efficiency without sacrificing performance.
- Quantization
  - Converting floating-point parameters to lower precision formats (e.g., integer or mixed-precision) to reduce memory footprint and computational overhead.

Optimizing Model Size and Performance

- Knowledge Distillation
  - Transferring knowledge from a large, complex model (teacher) to a smaller, more lightweight model (student) to achieve comparable performance with reduced computational resources.
- Architecture Design
  - Designing efficient model architectures that strike a balance between performance and resource utilization, such as using attention mechanisms.

Customizing Local Models

- Customizing local models empowers organizations to fine-tune AI solutions to their unique requirements.
- Leveraging open-source models and techniques like RAG, companies can tailor AI systems to specific use cases, enhance performance, and maintain data privacy.
- By optimizing models for local needs, organizations can achieve greater efficiency and effectiveness in their AI implementations.

Virtual Agents and Task Automation

- Virtual agents and task automation are reshaping industries by streamlining operations and enhancing customer experiences.
- From AI-powered chatbots that handle customer inquiries to intelligent automation systems that automate repetitive tasks, these solutions improve efficiency and productivity.
- By leveraging AI technologies, organizations can deliver faster responses, reduce human error, and improve overall service quality.

Regulatory Challenges and Compliance

- In the rapidly evolving landscape of AI development, navigating regulatory challenges and ensuring compliance is paramount.
  - Regulations like the EU's AI Act aim to address concerns related to data privacy, ethics, and accountability in AI usage.
- Understanding these regulations and implementing appropriate safeguards ensures that AI systems adhere to ethical standards and legal requirements, fostering trust and transparency.

Shadow AI and Security Issues

- The proliferation of shadow AI, or unauthorized AI usage in the workplace, presents significant security risks and compliance challenges.
    - From data breaches to algorithmic biases, the unauthorized deployment of AI technologies can have serious consequences.
- Implementing robust security measures, developing corporate AI policies, and fostering a culture of responsible AI usage are essential for mitigating risks and ensuring the safe and ethical deployment of AI solutions.

Take Home Exercise

Search for a company leveraging AI in a way that significantly improves or enhances their product or service. This can be any industry - retail, healthcare, finance, entertainment, etc. Avoid examples already covered in class.

- Here are some details to consider when making your selection:
  - What is the product or service?
  - How is AI used?
  - What are the benefits?
  - Are there any limitations?

Also read explore the PyTorch Documentation to get a better understanding of how to set up a model for the final project.