Unlocking the Power of Generative AI

Dr. BK Murthy

27-Jan-24

Artificial Intelligence, What is it?

- AI is defined as “**A system that shows behavior that could be interpreted as human intelligence**.” **-** Doug Rose

- AI thrives in an environment where there are **defined rules and patterns** that it can work with. This is where AI will seem the most “Intelligent”.

- If you have used **any of the following**, you have used AI:
  - Google Translate, Alexa, Siri, Chat GPT etc.

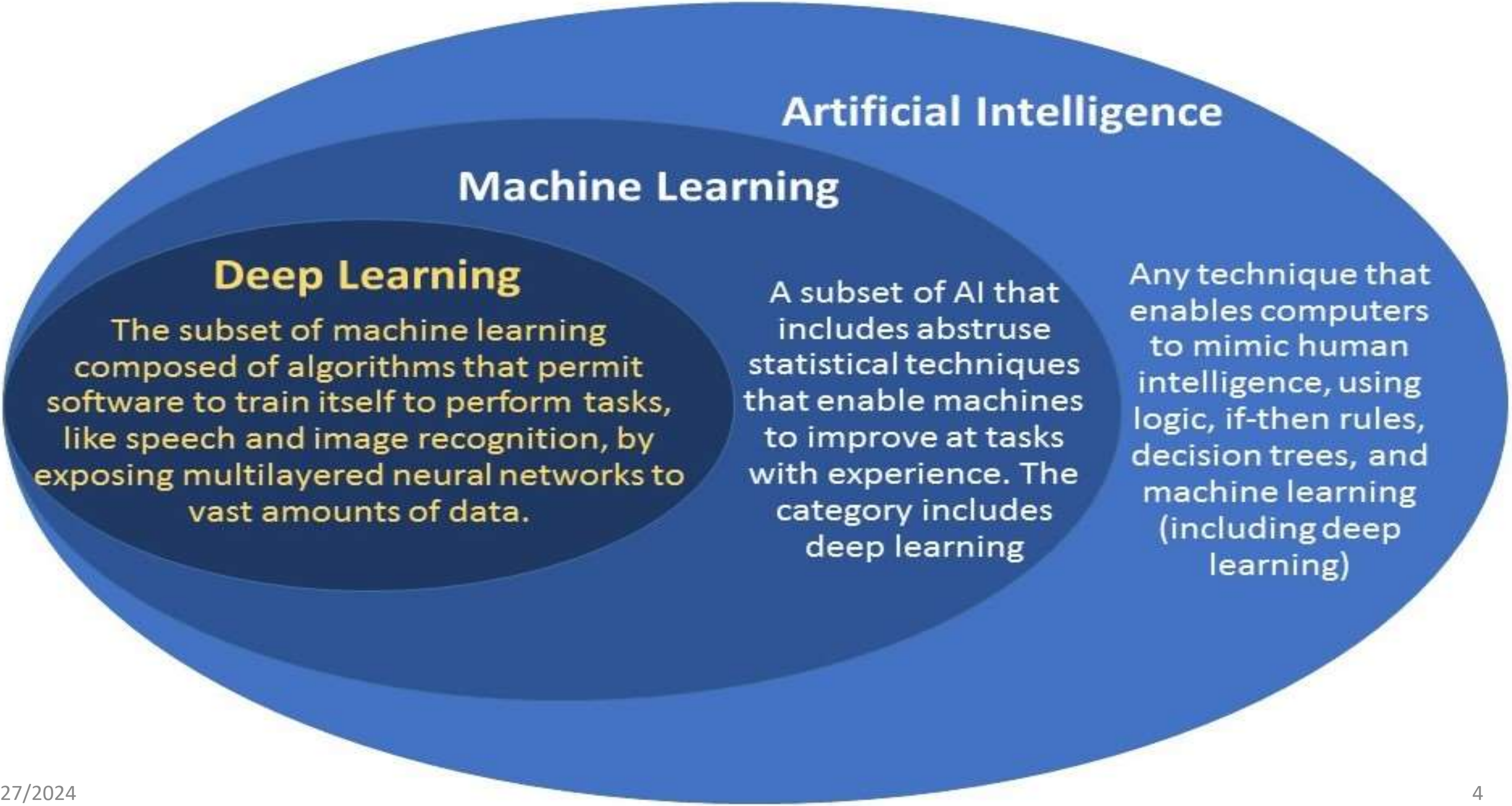
Artificial Intelligence (AI)

Artificial Intelligence deals with Design and development of computer systems that are able to perform tasks that Normally require human intelligence.

Key Technologies in AI

Machine Learning	Natural Lang. Processing	Speech Processing	Computer Vision	Robotics
Improve performance without specific instruction.	Work with text As humans do	Transcribe and generate speech with Accuracy	Identify objects, scenes, activities from images	Cognitive tech sensors, actuators.
Learn Big Data Analytics	Chatbot	Voicebots -	Facial recognition	IoT

Emerging Technologies – AI/ML



Artificial Intelligence

Machine Learning

Deep Learning

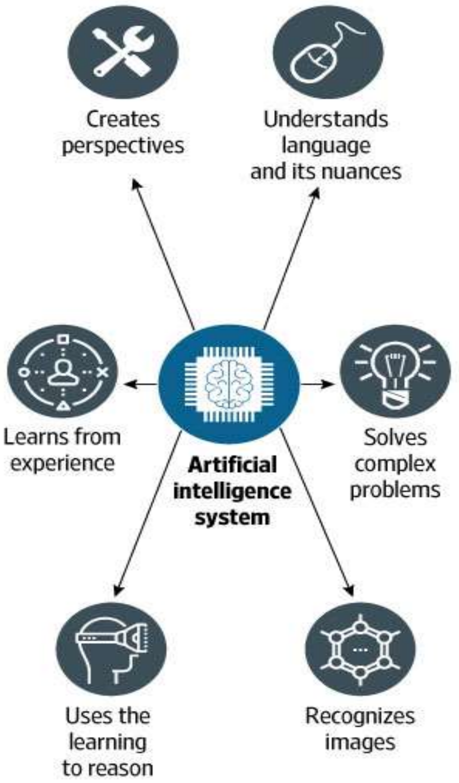
The subset of machine learning composed of algorithms that permit software to train itself to perform tasks, like speech and image recognition, by exposing multilayered neural networks to vast amounts of data.

A subset of AI that includes abstruse statistical techniques that enable machines to improve at tasks with experience. The category includes deep learning

Any technique that enables computers to mimic human intelligence, using logic, if-then rules, decision trees, and machine learning (including deep learning)

ARTIFICIAL INTELLIGENCE AND OPPORTUNITIES

How AI works



Artificial intelligence system

Creates perspectives

Understands language and its nuances

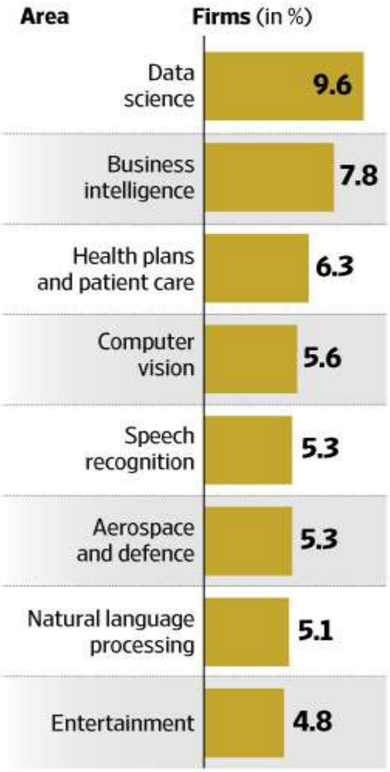
Solves complex problems

Recognizes images

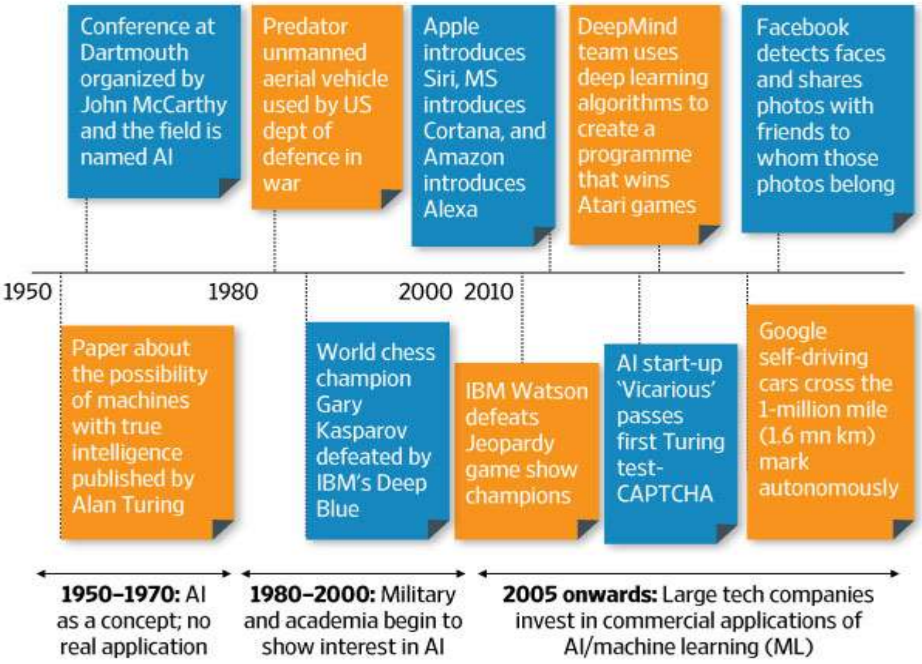
Uses the learning to reason

Learns from experience

Areas of focus for firms looking to exploit AI opportunities



Evolution of AI



1950 1980 2000 2010

Paper about the possibility of machines with true intelligence published by Alan Turing

Conference at Dartmouth organized by John McCarthy and the field is named AI

Predator unmanned aerial vehicle used by US dept of defence in war

World chess champion Gary Kasparov defeated by IBM's Deep Blue

IBM Watson defeats Jeopardy game show champions

Apple introduces Siri, MS introduces Cortana, and Amazon introduces Alexa

AI start-up 'Vicarious' passes first Turing test-CAPTCHA

DeepMind team uses deep learning algorithms to create a programme that wins Atari games

Google self-driving cars cross the 1-million mile (1.6 mn km) mark autonomously

Facebook detects faces and shares photos with friends to whom those photos belong

1950–1970: AI as a concept; no real application

1980–2000: Military and academia begin to show interest in AI

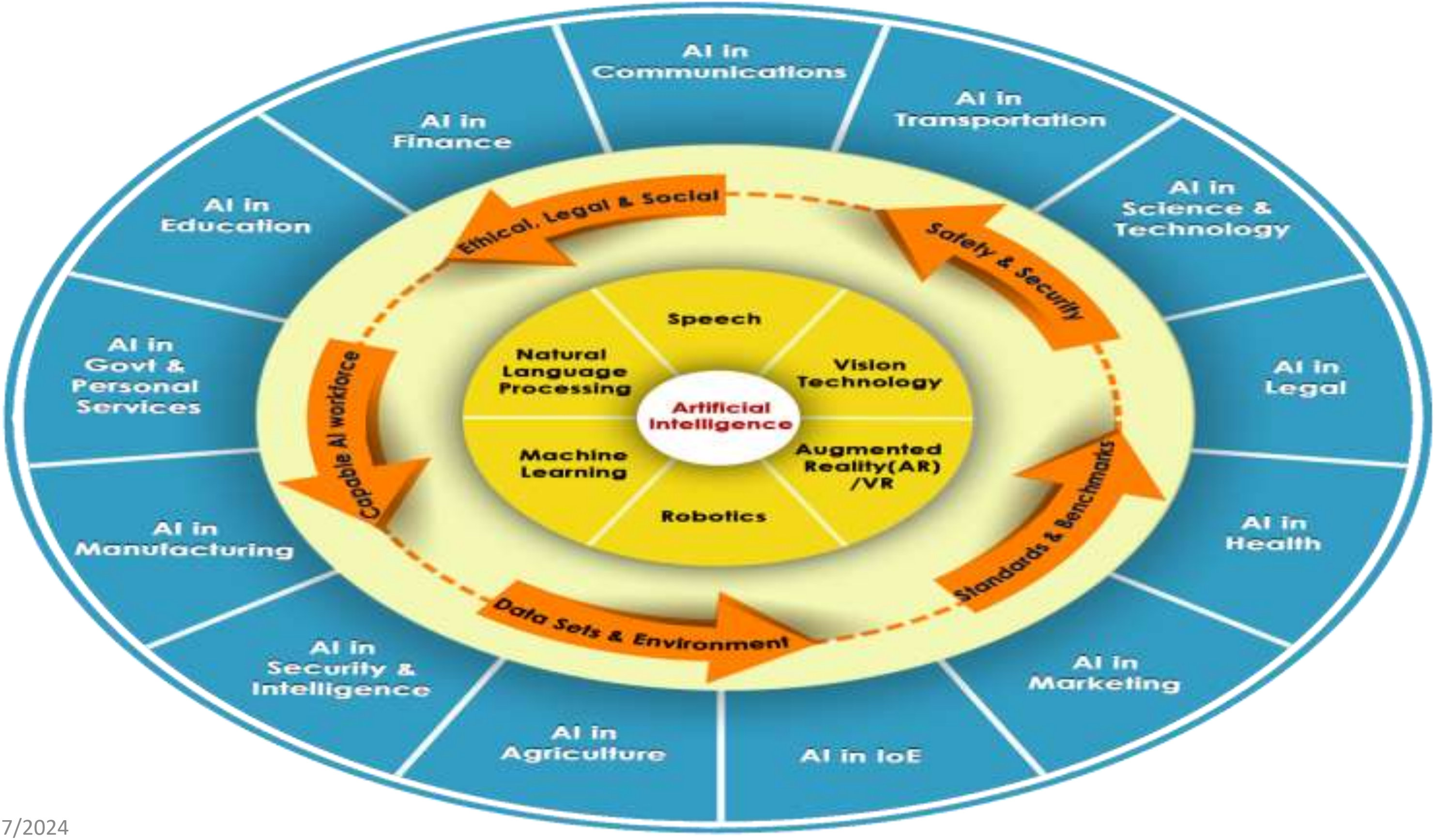
2005 onwards: Large tech companies invest in commercial applications of AI/machine learning (ML)

AI is currently on the crest of its third wave and given the immense interest in this field, it can be expected to continue at this state for a long time.

GRAPHIC: VIPUL SHARMA/MINT

Source: PwC Assocham report, AI and Robotics-2017

AI Core Domains and Applications



AI Impact

Economic, cultural, social, … endless disruption

Labour - McKinsey 58% of jobs automated

Martin Ford, Rise of the Robots

Elon Musk, artificial intelligence... existential threat

Strong AI Vs. Weak AI

- **Strong AI** is AI that acts exactly as a human would, think , the Terminator or Commander Data. They exhibit emotions, real creativity, and can even have a sense of purpose.

- **Weak AI** is AI that is confined to a narrow task, like when a system processes language into text or sorts all the pictures on your pc.

- Examples of Weak AI include: Siri, Cortana, Bing, Netflix, and even ChatGPT.

Introduction to Generative AI

- **Generative AI** refers to a type of artificial intelligence that has the ability to generate content that is, in many cases, **indistinguishable from content created by humans**. This AI can produce text, images, audio, or even video, often in response to a given input or prompt.

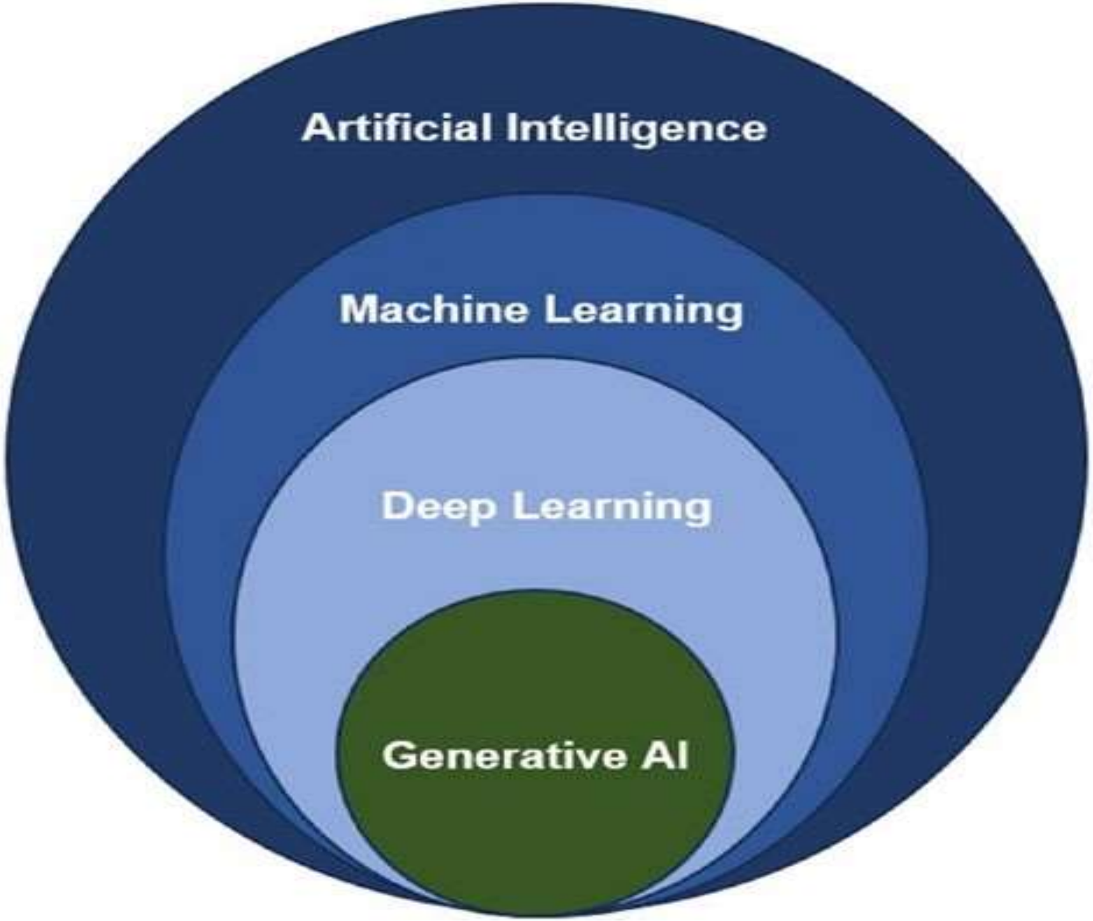
- **Generative AI** operates by **learning patterns and structures** from large datasets and **then using that knowledge to produce new content** that fits within those learned patterns. It's a type of machine learning where the AI model learns to understand and mimic the characteristics of the data it has been trained on.

Introduction

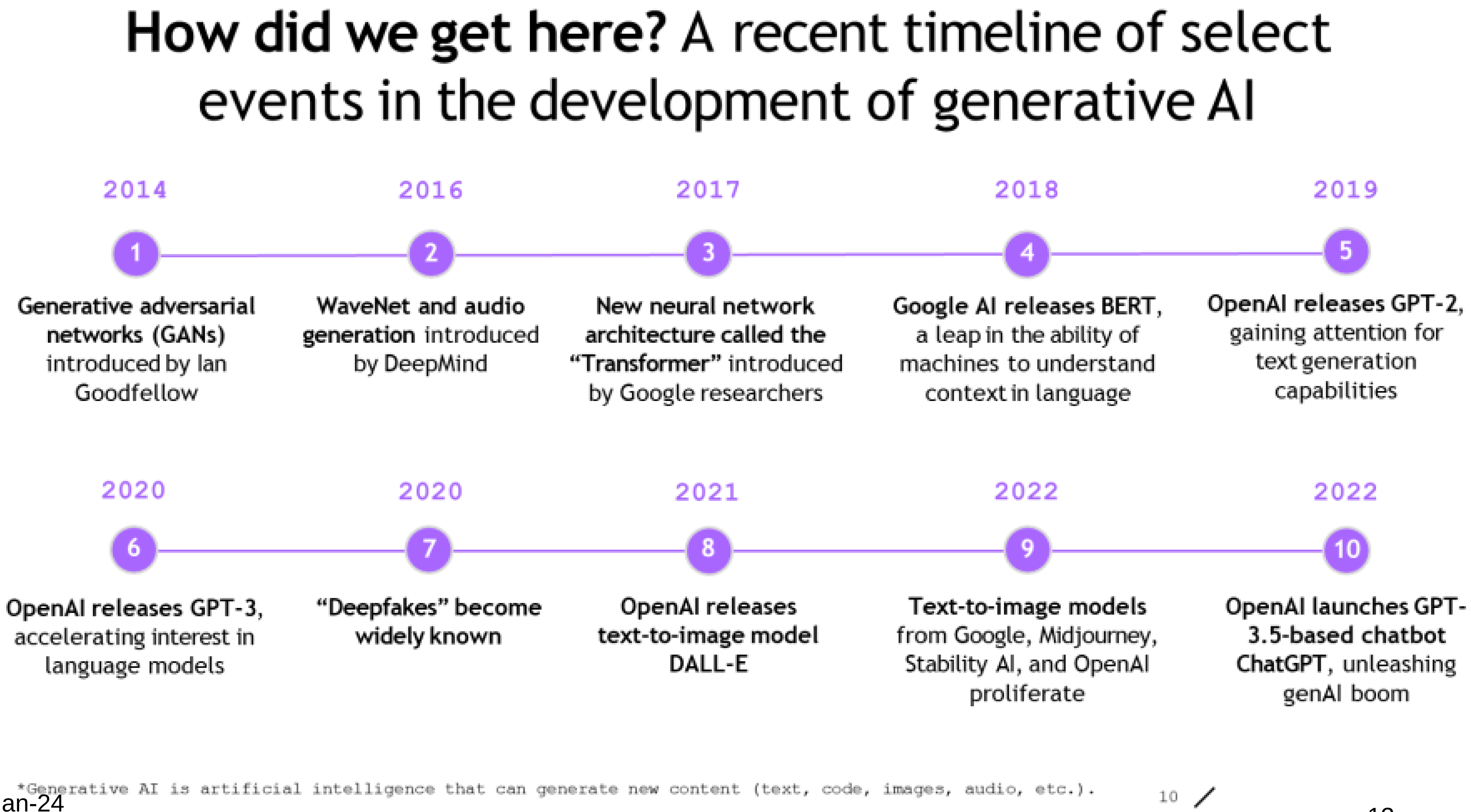
- **Generative AI** refers to a branch of artificial intelligence that focuses on creating models and algorithms capable of generating new, original content, such as images, text, music, and even videos.
- Unlike traditional AI models that are trained to perform specific tasks, generative AI models aim to learn and mimic patterns from existing data to generate new, unique outputs.
- The ability to generate realistic and convincing fake content can be misused for malicious purposes, such as creating deepfakes or spreading disinformation.
- As a result, there is ongoing research and development of techniques to detect and mitigate the potential negative impacts of generative AI.

Components of AI

- Artificial Intelligence (AI): It is the broader discipline of machine learning to perform tasks that would typically require human intelligence.
- Machine Learning (ML): A subset of AI, ML involves algorithms that allow computers to learn from data rather than being explicitly programmed to do so.
- Deep Learning (DL): A specialized subset of ML, deep learning involves neural networks with three or more layers that can analyze various factors of a dataset.
- Generative AI: An advanced subset of AI and DL, generative AI focuses on creating new and unique outputs. It goes beyond the scope of simply analyzing data to making new creations based on learned patterns.



Artificial Intelligence
Machine Learning
Deep Learning
Generative AI



How did we get here? A recent timeline of select events in the development of generative AI

2014

1

Generative adversarial networks (GANs) introduced by Ian Goodfellow

2016

2

WaveNet and audio generation introduced by DeepMind

2017

3

New neural network architecture called the "Transformer" introduced by Google researchers

2018

4

Google AI releases BERT, a leap in the ability of machines to understand context in language

2019

5

OpenAI releases GPT-2, gaining attention for text generation capabilities

2020

6

OpenAI releases GPT-3, accelerating interest in language models

2020

7

"Deepfakes" become widely known

2021

8

OpenAI releases text-to-image model DALL-E

2022

9

Text-to-image models from Google, Midjourney, Stability AI, and OpenAI proliferate

2022

10

OpenAI launches GPT-3.5-based chatbot ChatGPT, unleashing genAI boom

*Generative AI is artificial intelligence that can generate new content (text, code, images, audio, etc.).

Domains of Generative AI

- Text Generation
- Image Generation
- Audio Generation
- Video Generation

Text Generation

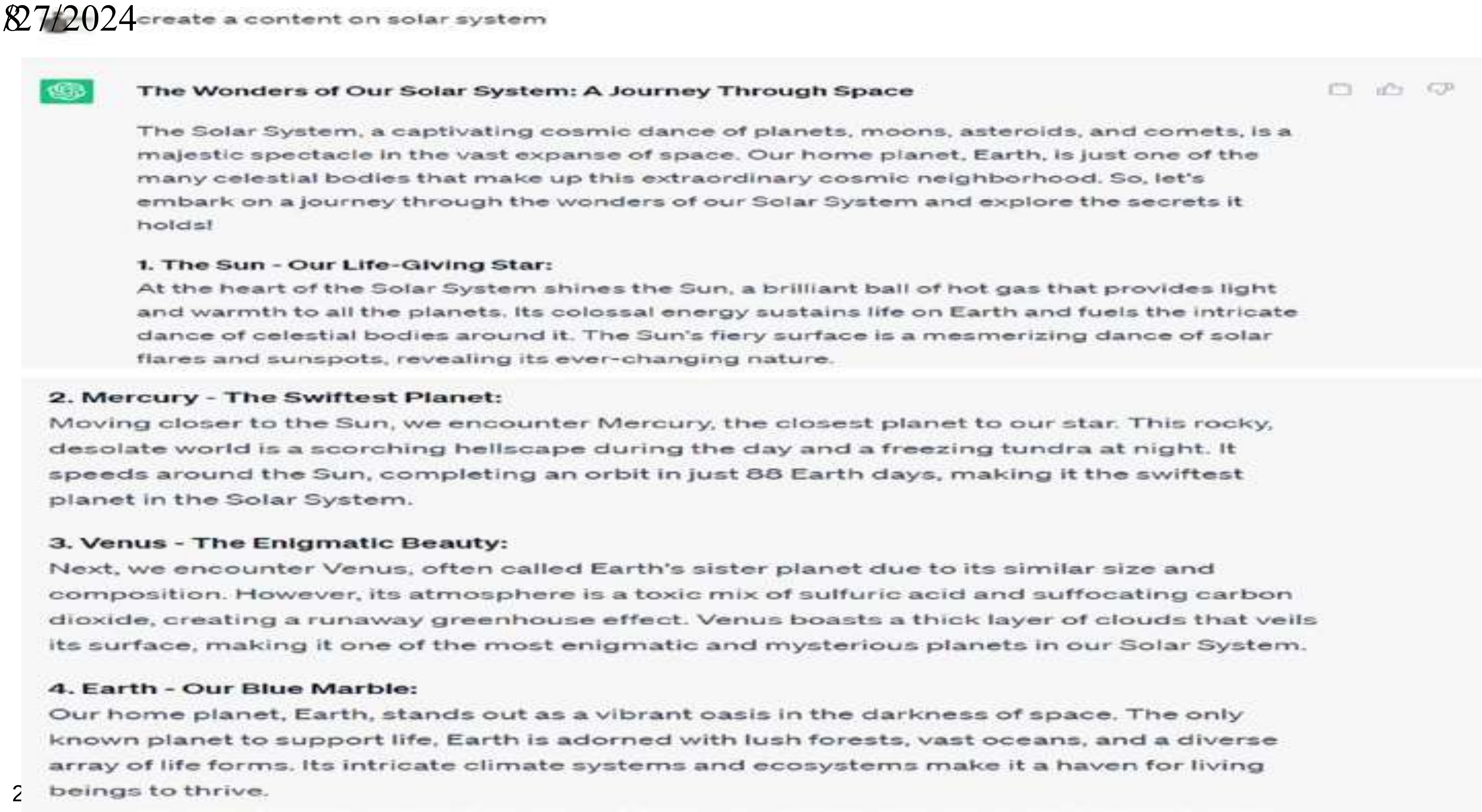
- What It Is: Text generation involves using AI models to create humanlike text based on input prompts.
- How It Works: Models like GPT-3 use Transformer architectures. They're pre-trained on vast text datasets to learn grammar, context, and semantics. Given a prompt, they predict the next word or phrase based on patterns they've learned.
- Applications:
  - Text generation is applied in content creation, chatbots, and code generation.
  - Businesses can use it for crafting blog posts, automating customer support responses, and even generating code snippets.
  - Strategic thinkers can harness it to quickly draft marketing copy or create personalized messages for customers.

Chat GPT

- It is a large language model–based chatbot developed by OpenAI and launched in November 2022.
- ChatGPT Uses Large Language Models created by Transformers (a variant of Neural Networks)
- ChatGPT is trained with reinforcement learning through human feedback and reward models that rank the best responses.
- This feedback helps augment ChatGPT with machine learning to improve future responses.

Applications

- Having conversations with users
- Answering questions
- Generating text
- Translating languages
- Writing different kinds of creative content
- ChatGPT can be accessed online at https://openai.com/blog/chatgpt



create a content on solar system

The Wonders of Our Solar System: A Journey Through Space

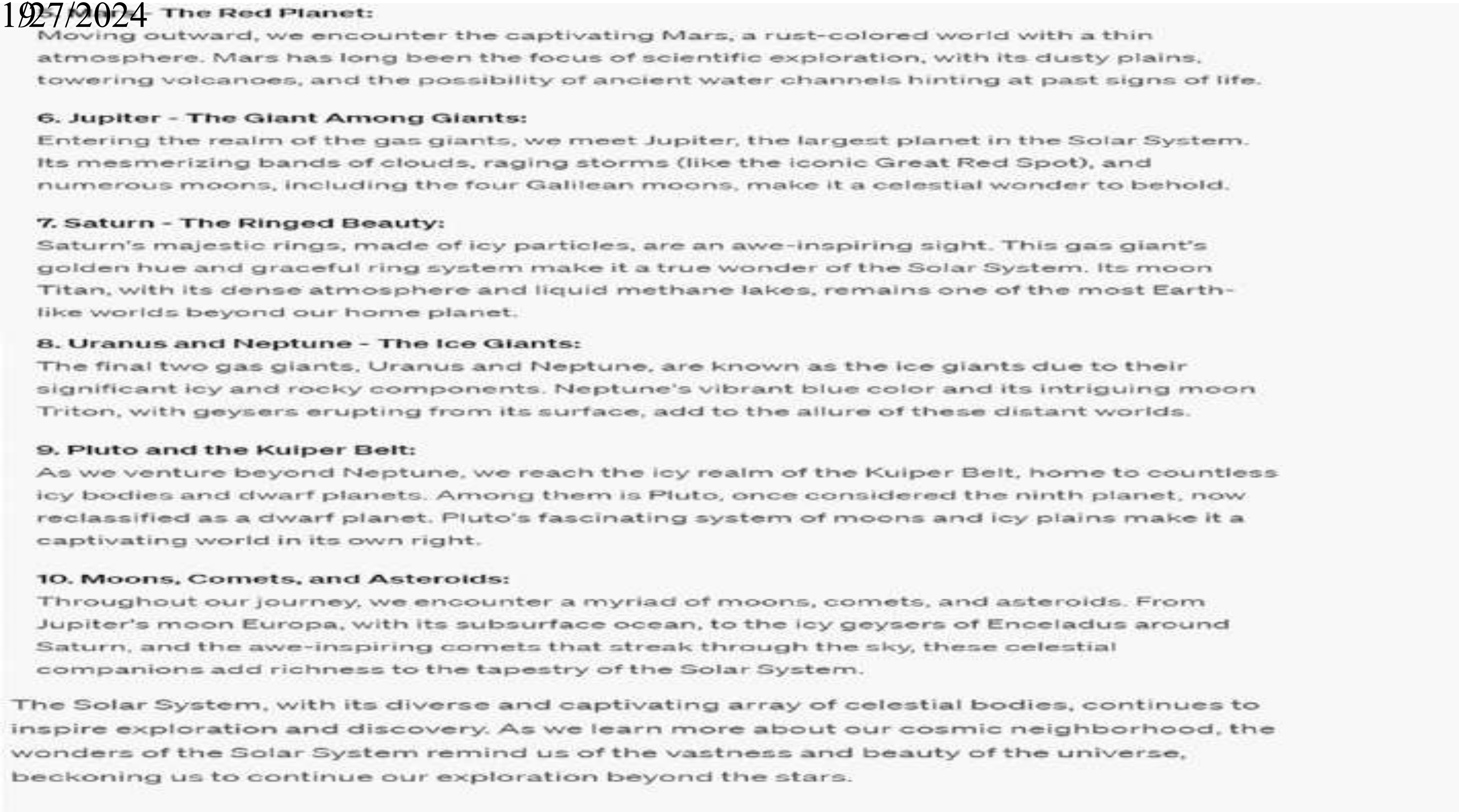
The Solar System, a captivating cosmic dance of planets, moons, asteroids, and comets, is a majestic spectacle in the vast expanse of space. Our home planet, Earth, is just one of the many celestial bodies that make up this extraordinary cosmic neighborhood. So, let's embark on a journey through the wonders of our Solar System and explore the secrets it holds!

1. The Sun - Our Life-Giving Star:
At the heart of the Solar System shines the Sun, a brilliant ball of hot gas that provides light and warmth to all the planets. Its colossal energy sustains life on Earth and fuels the intricate dance of celestial bodies around it. The Sun's fiery surface is a mesmerizing dance of solar flares and sunspots, revealing its ever-changing nature.

2. Mercury - The Swiftest Planet:
Moving closer to the Sun, we encounter Mercury, the closest planet to our star. This rocky, desolate world is a scorching hellscape during the day and a freezing tundra at night. It speeds around the Sun, completing an orbit in just 88 Earth days, making it the swiftest planet in the Solar System.

3. Venus - The Enigmatic Beauty:
Next, we encounter Venus, often called Earth's sister planet due to its similar size and composition. However, its atmosphere is a toxic mix of sulfuric acid and suffocating carbon dioxide, creating a runaway greenhouse effect. Venus boasts a thick layer of clouds that veils its surface, making it one of the most enigmatic and mysterious planets in our Solar System.

4. Earth - Our Blue Marble:
Our home planet, Earth, stands out as a vibrant oasis in the darkness of space. The only known planet to support life, Earth is adorned with lush forests, vast oceans, and a diverse array of life forms. Its intricate climate systems and ecosystems make it a haven for living beings to thrive.



5. Mars - The Red Planet:
Moving outward, we encounter the captivating Mars, a rust-colored world with a thin atmosphere. Mars has long been the focus of scientific exploration, with its dusty plains, towering volcanoes, and the possibility of ancient water channels hinting at past signs of life.

6. Jupiter - The Giant Among Giants:
Entering the realm of the gas giants, we meet Jupiter, the largest planet in the Solar System. Its mesmerizing bands of clouds, raging storms (like the iconic Great Red Spot), and numerous moons, including the four Galilean moons, make it a celestial wonder to behold.

7. Saturn - The Ringed Beauty:
Saturn's majestic rings, made of icy particles, are an awe-inspiring sight. This gas giant's golden hue and graceful ring system make it a true wonder of the Solar System. Its moon Titan, with its dense atmosphere and liquid methane lakes, remains one of the most Earth-like worlds beyond our home planet.

8. Uranus and Neptune - The Ice Giants:
The final two gas giants, Uranus and Neptune, are known as the ice giants due to their significant icy and rocky components. Neptune's vibrant blue color and its intriguing moon Triton, with geysers erupting from its surface, add to the allure of these distant worlds.

9. Pluto and the Kuiper Belt:
As we venture beyond Neptune, we reach the icy realm of the Kuiper Belt, home to countless icy bodies and dwarf planets. Among them is Pluto, once considered the ninth planet, now reclassified as a dwarf planet. Pluto's fascinating system of moons and icy plains make it a captivating world in its own right.

10. Moons, Comets, and Asteroids:
Throughout our journey, we encounter a myriad of moons, comets, and asteroids. From Jupiter's moon Europa, with its subsurface ocean, to the icy geysers of Enceladus around Saturn, and the awe-inspiring comets that streak through the sky, these celestial companions add richness to the tapestry of the Solar System.

The Solar System, with its diverse and captivating array of celestial bodies, continues to inspire exploration and discovery. As we learn more about our cosmic neighborhood, the wonders of the Solar System remind us of the vastness and beauty of the universe, beckoning us to continue our exploration beyond the stars.

Image Generation

What It Is:

- Image generation involves using various deep learning models to create images that look real.

How It Works:

- GANs consist of a generator (creates images) and a discriminator (determines real vs. fake).
- They compete in a feedback loop, with the generator getting better at producing images that the discriminator can’t distinguish from real ones.

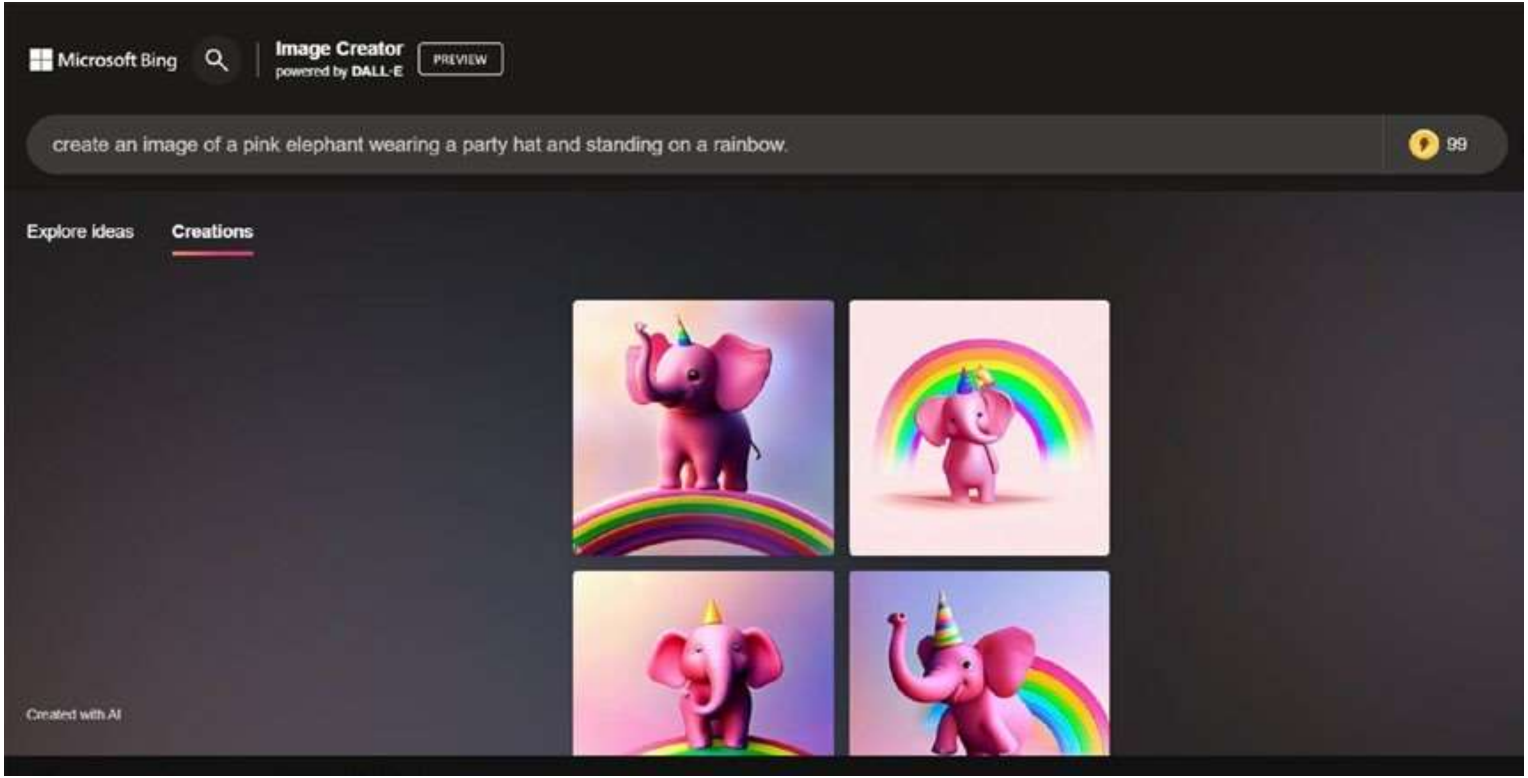
Applications:

- These models are used in art, design, and product visualization. Businesses can generate product mock-ups for advertising, create unique artwork for branding, or even generate faces for diverse marketing materials.

- Microsoft Bing Image Creator is a generative AI tool that uses artificial intelligence to create images based on your text descriptions.

  www.bing.com/images/create/
- To use Bing Image Creator, you simply type a description of the image you want to create into the text box.
- Example**:** “Create an image of a pink elephant wearing a party hat and standing on a rainbow.”
- Bing Image Creator will then generate an image based on your description.

Binge Created Pink Elephant Image



create an image of a pink elephant wearing a party hat and standing on a rainbow.

Audio Generation

What It Is:

- Audio generation involves AI creating music, sounds, or even humanlike voices.

How It Works:

- Models like WaveGAN analyze and mimic audio waveforms. Text-to-speech models like Tacotron 2 use input text to generate speech. They’re trained on large datasets to capture nuances of sound.

Applications:

- AI-generated music can be used in ads, videos, or as background tracks.
- Brands can create catchy jingles or custom sound effects for marketing campaigns.
- Text-to-speech technology can automate voiceovers for ads or customer service interactions.
- Strategically, businesses can use AI-generated audio to enhance brand recognition and storytelling.

Video Generation

What It Is:

- Video generation involves AI creating videos, often by combining existing visuals or completing missing parts.

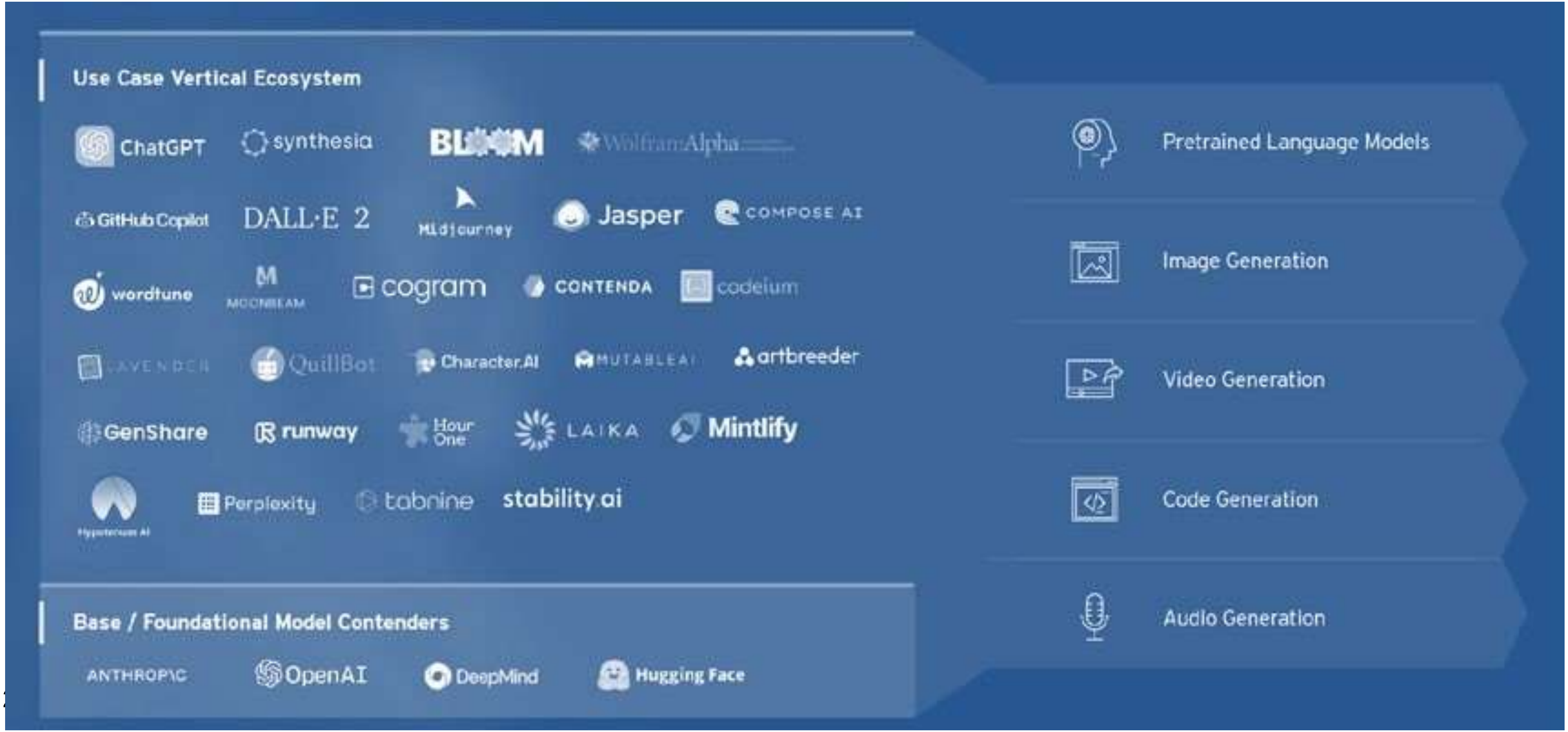
How It Works:

- Video generation is complex due to the temporal nature of videos. Some models use text descriptions to generate scenes, while others predict missing frames in videos.

Applications:

- AI-generated videos can be used in personalized messages, dynamic ads, or even content marketing.
- Brands can craft unique video advertisements tailored to specific customer segments.
- Thoughtful application can lead to efficient video content creation that adapts to marketing trends.

Generative AI Platforms



Generative AI Platforms contd..

- OpenAI: OpenAI is a generative AI research company that was founded by Elon Musk, Sam Altman, and others.
- OpenAI has developed some of the most advanced generative AI models in the world, including GPT-4 and DALL-E 2.
    - GPT-4: GPT-4 is a large language model that can generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way.
    - DALL-E 2: DALL-E 2 is a generative AI model that can create realistic images from text descriptions.

DeepMind:

- DeepMind is a British artificial intelligence company that was acquired by Google in 2014.
- DeepMind has developed several generative AI models, including
  - AlphaFold, which can predict the structure of proteins, and
  - Gato, which can perform a variety of tasks, including playing Atari games, controlling robotic arms, and writing different kinds of creative content.

Anthropic:

- Anthropic is a company that is developing generative AI models for use in a variety of industries, including healthcare, finance, and manufacturing.
- Anthropic’s models are trained on massive datasets of real-world data, which allows them to generate realistic and accurate outputs.

Synthesia:

- Synthesia is a company that specializes in creating realistic synthetic media, such as videos and audio recordings.
- Synthesia’s technology can be used to create avatars that can speak, gesture, and even lip-sync to any audio input.
  - RealSpeaker: RealSpeaker is a generative AI model that can be used to create realistic synthetic voices.
  - Natural Video: Natural Video is a generative AI model that can be used to create realistic synthetic videos.

RunwayML:

- RunwayML is a platform that makes it easy for businesses to build and deploy generative AI models.
- RunwayML provides a variety of tools and resources to help businesses collect data, train models, and evaluate results.
  - Runway Studio: Runway Studio is a cloud-based platform that allows businesses to build and deploy generative AI models without any coding experience.
  - Runway API: The Runway API is a set of APIs that allow businesses to integrate generative AI into their applications.
- Midjourney: Midjourney is a generative AI model that can be used to create realistic images, videos, and text.
- Midjourney is still underdevelopment, but it has already been used to create some impressive results.

Tools, Platforms, and Software

- **Midjourney/Dall-E3** –text to art
- **Wisdolia** – plugin, generate flash cards for any website, video, or PDF you are on.
- **Microsoft 365 copilot** – brings AI across the entire Microsoft office suite
- **Eleven Labs** – voice recognition. You speak to it, then you can feed it scripts and it will read them in your voice and cadence.
- **Mixo/Sitekick** – type a product idea and it creates a full website.
- **Tome** – makes presentations from simple prompts.
- **Tableau's Ask Data** – ask questions, receive data visualizations as responses.

Generative AI Applications

- Content Creation:
  - Text Generation: Automating blog posts, social media updates, and articles, eassays, letters email responses, complaint notices etc.
  - Image Generation: Creating custom visuals for marketing campaigns and advertisements.
  - Video Generation: Crafting personalized video messages and dynamic ads.
- Design and Creativity:
  - Art Generation: Creating unique artworks, illustrations, and designs.
  - Fashion Design: Designing clothing patterns and accessories.
  - Product Design: Generating prototypes and mock-ups.

Generative AI Applications contd..

- Entertainment and Media:
  - Music Composition: Creating original music tracks and soundscapes.
  - Film and Animation: Designing characters, scenes, and animations.
  - Storytelling: Developing interactive narratives and plotlines.
- Marketing and Advertising:
  - Personalization: Crafting tailored messages and recommendations for customers.
  - Branding: Designing logos, packaging, and visual identity elements.
  - Ad Campaigns: Developing dynamic and engaging advertisements.

- Gaming:
  - World Building: Generating game environments, terrains, and landscapes.
  - Character Design: Creating diverse and unique in-game characters.
  - Procedural Content: Generating levels, quests, and challenges.
- Healthcare and Medicine:
  - Drug Discovery: Designing new molecules and compounds.
  - Medical Imaging: Enhancing and reconstructing medical images.
  - Personalized Medicine: Tailoring treatment plans based on patient data.

- Language Translation:
  - Real-time Translation: Enabling instant translation of spoken or written language.
  - Subtitling and Localization: Automatically generating subtitles for videos.
- Customer Service:
  - Chatbots: Creating conversational agents for customer support.
  - Voice Assistants: Providing voice-based assistance for inquiries and tasks.

- Education and Training:
  - Interactive Learning: Developing adaptive learning materials.
  - Simulations: Creating realistic training scenarios and simulations.
- Architecture and Design:
  - Building Design: Generating architectural layouts and designs.
  - Urban Planning: Designing cityscapes and urban layouts.

Generative AI Limitations

- Quality and Coherence : Generative AI can sometimes produce content that is factually incorrect or incoherent.
- Lack of Understanding: Generative models don't have true understanding of the text they generate.
- They generate responses based on statistical patterns rather than comprehension, which means they can't answer questions that require deep understanding or common-sense reasoning.
- Biases: Generative AI can inadvertently perpetuate biases present in the training data.

Generative AI Limitations

- **Safety and Privacy**: In some cases, generative AI can generate harmful or inappropriate content. Ensuring the safety and ethical use of AI-generated text is a significant concern.
- **Inconsistency**: The same prompt given to a generative model may produce different responses at different times. While this can be useful for creativity, it can also result in inconsistent or contradictory answers.
- **Overgeneration**: Generative models can be verbose and tend to over generate content.

Generative AI Limitations

- **Data Dependency**: The quality of the generated text depends on the quality and diversity of the training data. Limited or biased training data can result in poor performance.
- ChatGPT3.5 for example only has data up to September 2021.

- **Prompt Sensitivity**: The way a prompt is framed can significantly impact the output.
- Crafting effective prompts requires skill and experimentation.

Identifying Opportunities for AI

- **Nature of the Task:** What are you trying to generate?
- **Complexity of the Task:** Does it need to be broken into segments?
- **Data Availability:** How recent/prevalent is the data for what you are trying to do?
- **Ethical Considerations:** Use ethical guidelines to avoid harmful or biased content.
- **Human Review/Monitoring:** Human oversight is needed to ensure no errors or biases are present.
- **Scalability:** Assess if the task can be handled efficiently with available computational resources.

Real-World Applications in Universities

(Students)

Study and Homework Assistance:

- Generating explanations and solutions for homework problems.
- Providing study tips and summaries of course materials.
- Offering virtual tutors for a wide range of subjects.
- Explaining complex concepts and answering questions.
- Creating exercises, quizzes, and pronunciation guides.
- Helping students improve their essays, reports, and creative writing.

Personalized Learning Plans:

- Analyzing students' performance data to recommend customized study plans.
- Suggesting additional reading materials and resources.

Career Guidance:

- Providing advice on choosing majors and career paths based on students' interests and skills.
- Assisting in resume and cover letter writing.

Research Assistance:

- Assisting in gathering preliminary research data and suggesting relevant sources.
- Generating citations and bibliographies.

Language Translation:

- Translating foreign language texts and documents for international students.
- Supporting international exchange programs.

Time Management and Organization:

- Creating personalized schedules and reminders for classes and assignments.
- Offering productivity tips and techniques.

Real-World Applications in Universities

(Staff)

Content Generation:

- Automating the creation of course materials, lecture notes, and assessments.
- Generating content for university publications and marketing materials

Administrative Support:

- Assisting in scheduling meetings, managing emails, and handling routine administrative tasks.
- Answering frequently asked questions for staff and faculty.

Research Assistance:

- Analyzing and summarizing research papers and articles.
- Assisting in data analysis and visualization.

Admissions and Enrollment:

- Managing inquiries from prospective students.
- Automating admissions and enrollment processes.

Emergency Response and Communications:

- Providing automated alerts and communication during campus emergencies.
- Offering guidance on emergency protocols and procedures.

Student Support Services:

- Providing automated responses to student inquiries related to registration, financial aid, and campus resources.
- Offering career counseling and internship recommendations

Website and Social Media Management:

- Generating content for university websites, blogs, and social media platforms.
- Monitoring and responding to online engagement.

Grading and Assessment:

- Assisting in grading assignments and exams.
- Analyzing student performance data to identify areas for improvement.

Library and Information Services:

- Assisting in information retrieval and research assistance for both faculty and students.
- Automating library cataloging and resource recommendations.

Case Study #1: Admissions

- Elite University, a renowned higher education institution, has a highly competitive college admissions department that receives thousands of applications each year.
- To improve their efficiency and provide a better experience for applicants, Elite University decided to implement generative AI solutions.
- This case study details how the college admissions department leveraged generative AI to enhance their performance.

Challenges:

1. **Application Processing**: Reviewing and processing a large volume of applications was a time-consuming and labor-intensive task, often leading to bottlenecks and delays.
2. **Essay Assessment**: Assessing essays and personal statements for admissions required significant manual effort.

   It was difficult to ensure consistency and fairness in evaluations.
3. **Applicant Support**: Applicants often had questions about the application process, requirements, and deadlines.

   Providing timely and accurate responses to these inquiries was a challenge.

Solution:

- **Automated Application Screening**:
  - Integrated AI algorithms to automatically screen and categorize applications based on predefined criteria. This allowed the admissions team to prioritize applications that met minimum requirements.
  - The generative AI system learned from historical admissions data to continuously refine its screening criteria.
- **AI-Enhanced Essay Assessment**:
  - Employed generative AI to assist in the assessment of essays and personal statements. The system provided initial evaluations, highlighting key strengths and areas for improvement.
  - Admissions officers could use the AI-generated assessments as a starting point, saving time while ensuring a standardized review process.

Solution:

- **AI-Powered Applicant Support**:
  - Implemented AI-powered chatbots on the university's admissions website and application portal. These chatbots answered applicant inquiries regarding deadlines, requirements, and procedures.
  - Chatbots were trained using frequently asked questions and were designed to provide accurate and up-to-date information.

Results:

- **Faster Application Processing**:
  - Automated application screening **reduced the time required to process applications**, ensuring that qualified applicants progressed to the next stages more quickly.
- **Improved Essay Assessment**:
  - AI-assisted essay assessments provided consistent evaluations, reducing bias and ensuring fairness in the admissions process.
  - **Admissions officers had more time to focus on nuanced evaluations of applicant essays**.

Results cntd…

- **Enhanced Applicant Experience**:
  - AI-powered chatbots provided **quick and accurate responses** to applicant inquiries, improving the overall applicant experience.
- **Increased Efficiency**:
  - Admissions staff experienced increased efficiency as routine tasks were automated, **allowing them to allocate more time to strategic decision-making.**
- **Data-Driven Insights**:
  - The generative AI system collected and analyzed data on applicant behavior, providing valuable insights into the admissions process and applicant preferences.

Prompt Engineering

- **Prompt engineering is the process of designing and crafting input prompts or queries to generative AI models to elicit desired outputs or responses.**
- The choice of words, format, and context in the prompt can significantly influence the generated content.
- How to structure prompts for desired outputs:

- Be Clear and Specific
- Specify the Format
- Add Context
- Use Examples
- Control the Tone

- Ask the Model to Think Step by Step
- Use Keywords
- Provide Constraints
- Experiment
- Iterate and Refine

Prompt Engineering Examples

- **Task: Summarize a Report**
  - Ineffective Prompt: "Summarize this report."
  - Effective Prompt: "Provide a concise summary of the key findings and overarching messages of the GLBA Audit Findings: [paste report here]. "
- **Task: Creative Writing**
  - Ineffective Prompt: "Write a story."
  - Effective Prompt: "Create an engaging short story about a time traveler who finds themselves in a parallel universe where gravity behaves differently."
- **Task: Language Translation**
  - Ineffective Prompt: "Translate this sentence."
  - Effective Prompt: "Translate the following English sentence into French: 'The quick brown fox jumps over the lazy dog.'"

Activity

We are going to test chatGPT's effectiveness on some real-world scenarios.

- Warm up: **Consider what aspects of your work could be enhanced or made easier by using Generative AI.**

- **Questions to ask yourself:**
  - what tasks are repetitive in my job?
  - what kind of writing do you do that could be done by AI?
  - what projects could you use help organizing or starting?

Activity

Now that we have thought of a few things that we can use generative AI for in our jobs, let’s practice! **Try to get ChatGPT to perform some of the tasks you have thought of.**

Some examples if you need somewhere to get started:

- Write a newsletter to students about essential student services here at the University
- Write a social media post about the DIT University
- Write a project plan/outline
- Draft an email you normally have trouble starting

Information Security, Privacy, and Ethics

- **Do tools and platforms like ChatGPT present an inherent security risk?**
  - From their Terms of Service (TOS): “Use of Content to Improve Services: We do not use Content that you provide to or receive from our API (“API Content”) to develop or improve our Services.
  - **We may use Content from Services other than our API (“Non-API Content”) to help develop and improve our Services**.”
  - OpenAI Recommends using fake names or pseudonyms when interacting with ChatGPT, and to avoid public wi-fi, instead using secured private networks.

Information Security, Privacy, and Ethics contd…

- **Not all platforms follow the same or even similar guidelines**
  - "Copilot seamlessly integrates into Microsoft 365, **inheriting your organization's security, compliance, and privacy policies, It utilizes advanced encryption, access control, and permissions to prevent data leakage and maintain compliance with security and privacy policies**.
  - Microsoft Copilot places a high emphasis on data security and privacy within Microsoft 365." - Microsoft

Ethical Concerns

- **Bias and Fairness**: AI systems can inherit biases from the data they are trained on, potentially leading to **discrimination in areas like admissions, hiring, or grading**.
- **Privacy:** AI may **process and store sensitive student or faculty data**, raising concerns about data security and privacy violations.
- **Transparency:** The opacity of some AI algorithms makes it difficult to understand how decisions are reached. **This lack of transparency can raise ethical questions about accountability and trust**.

Ethical Concerns contd…

- **Accountability**: It can be challenging to assign responsibility when AI is used in decision-making processes. **Determining who is accountable for AI-related outcomes or errors is an important ethical consideration.**
- **Data Quality**: **Garbage in, garbage out**: If AI systems are fed with poor-quality or biased data, the ethical integrity of the resulting decisions is compromised.
- **Consent**: Collecting and using personal data for AI applications should involve informed consent. **Universities must be transparent about data usage and give individuals the option to opt in or out.**

Guidelines for Leveraging Generative AI

- **Understand the Technology**
  - Ensure that you and your team have a deep understanding of how generative AI works, its capabilities, and its limitations. This understanding is crucial for responsible use.
- **Data Ethics**
  - Use high-quality and diverse training data that is free from bias and sensitive information. Be aware of the potential biases in your training data and take steps to mitigate them.
- **Human Oversight**
  - Maintain human oversight and control over AI systems. Fact-check your data and avoid plagiarism.

Guidelines for Leveraging Generative AI contd..

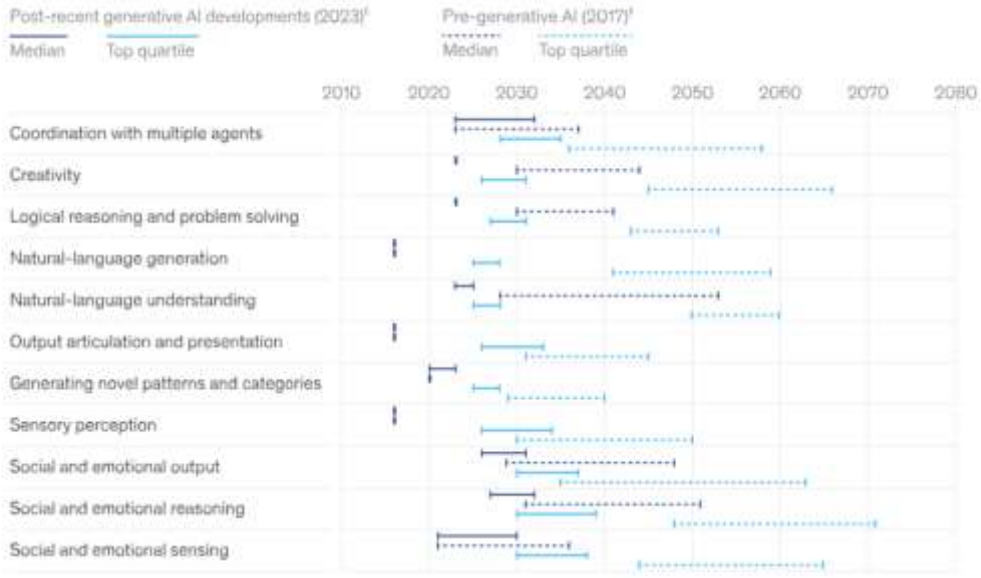
- **Accountability**:
  - Clearly define roles and responsibilities for AI development and deployment. Ensure accountability for the outcomes of AI systems, both positive and negative.
- **Education and Training**:
  - Provide training and guidelines to staff or users interacting with AI systems to promote responsible usage and ethical considerations.
- **Continual Monitoring and Evaluation**:
  - Continuously monitor the performance and impact of AI systems after deployment. Be prepared to make adjustments or take corrective actions as needed.

Future Trends

- For most of the technical capabilities shown in this chart, **gen AI will perform at a median level of human performance by the end of this decade**.
- Its performance will compete with the top 25 percent of people completing any and all of these tasks before 2040**.**
- **In some cases, that’s 40 years faster than experts previously thought**.

Due to generative AI, experts assess that technology could achieve human-level performance in some capabilities sooner than previously thought.

Estimated range for technology to achieve human-level performance, by technical capability

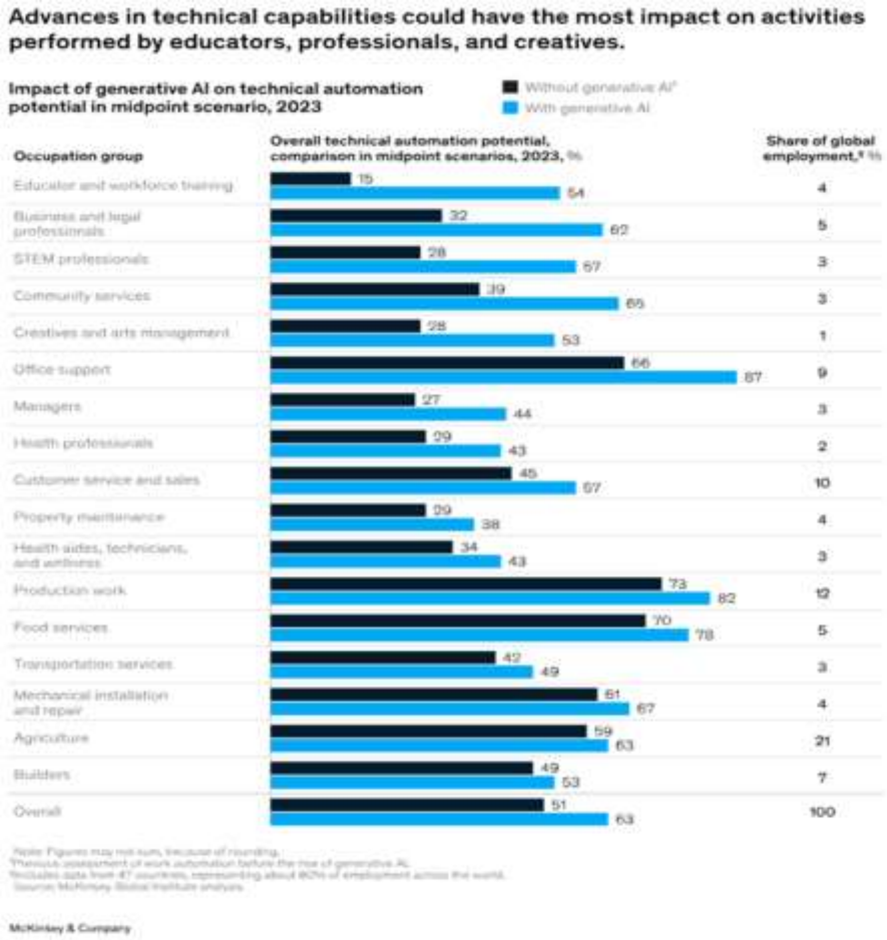


[1] Comparison made on the business-related tasks required from human workers.
Source: McKinsey Global Institute occupation database; McKinsey analysis

McKinsey & Company

Future Trends

- Previous waves of automation technology mostly affected physical work activities, but **gen AI is likely to have the biggest impact on knowledge work—especially activities involving decision making and collaboration.**
- Professionals in fields such as education, law, technology, and the arts are likely to see parts of their jobs automated sooner than previously expected.
- This is because of generative AI's ability to predict patterns in natural language and use it dynamically.



Advances in technical capabilities could have the most impact on activities performed by educators, professionals, and creatives.

Impact of generative AI on technical automation potential in midpoint scenario, 2023

Note: Figures may not sum, because of rounding.
¹Previous assessment of work automation before the rise of generative AI.
²Includes data from 47 countries, representing about 80% of employment across the world.
Source: McKinsey Global Institute analysis

McKinsey & Company

Summary

- Generative AI, a rapidly evolving domain in artificial intelligence that specializes in creating new, unique content such as text, images, audio, and videos.
- Built upon advancements in deep learning and natural language processing (NLP), these models have various applications, including content creation, design, entertainment, healthcare, and customer service.
- Generative AI also brings ethical concerns, particularly in creating deepfakes or spreading disinformation.
- Various Applications of generative AI for text, image, audio, and video generation—detailing how they work and their practical applications.
- Discussed some of the key players in the industry, like OpenAI, DeepMind, and Synthesia, among others. Lastly, it outlines a wide array of applications across various industries.

What can Artificial Intelligence generate?

- Large language models such as ChatGPT can generate text in multiple languages and styles. They can also interpret text and images.
- Generative programs such as DALL.E2 and Midjourney generate images from text instructions. Others can generate video, audio, and code.
- AI will be increasingly blended into everyday software as a task assistant.
- ChatGPT has an increasing list of plugins that extend the natural language model to connect to services, data, and calculations.

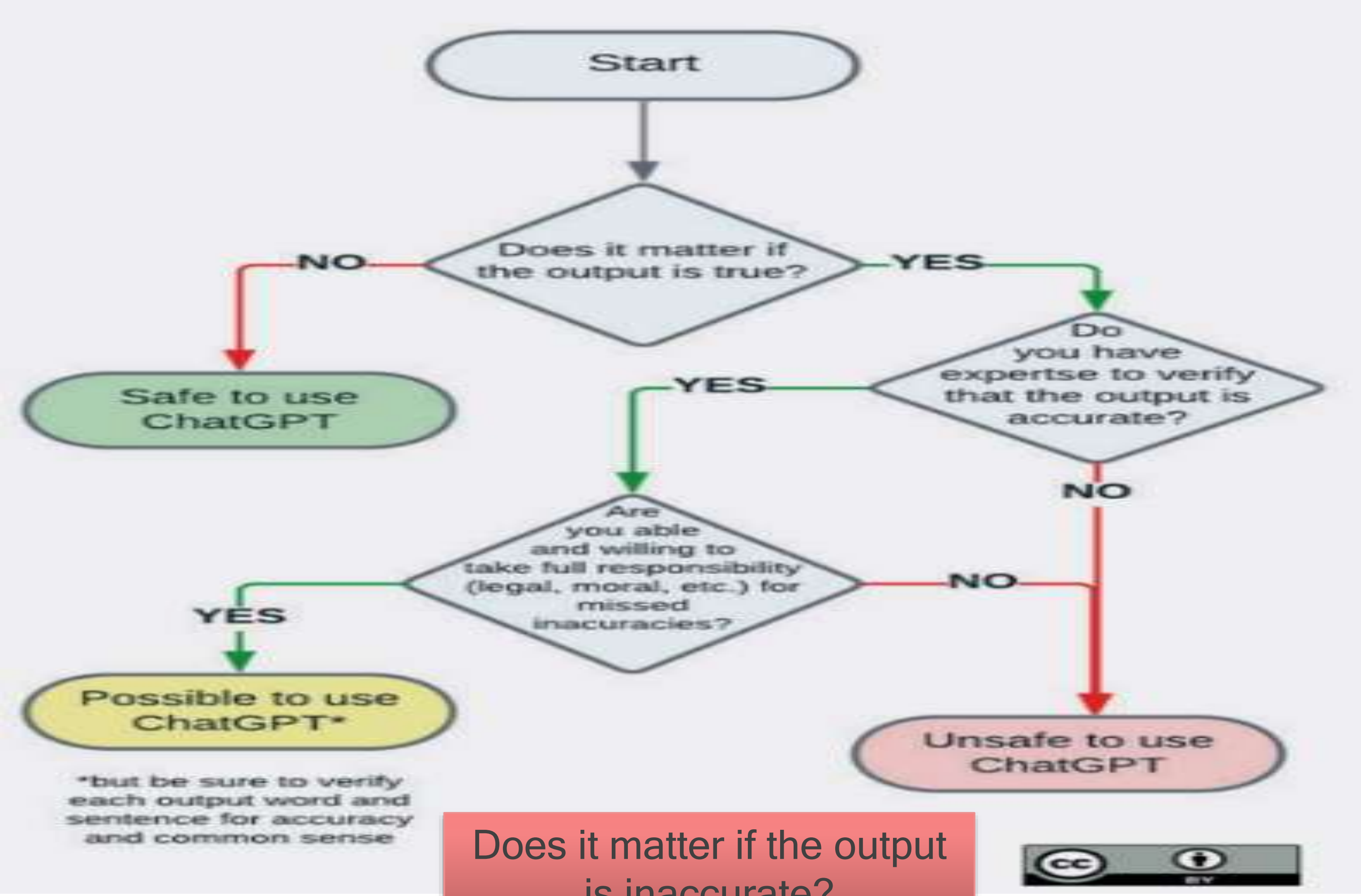
What are large language models (LLMs)?

- Large language models (LLMs) are a type of advanced artificial intelligence system. ChatGPT is an example of an LLM that allows people to interact with a computer in a more natural and conversational way.
- GPT stands for “Generative Pre-trained Transformer” and is the name given to a family of natural language models developed by OpenAI. There are other large language models such as Bard, LLaMA, and Claude.
- These tools are known as **generative AI** because of their ability to produce seemingly original results.
- They are trained on large text datasets to learn to predict the next word in a sentence and, from that, generate coherent and compelling responses. GPT-3 is trained on 300 billion words.

What do you think are some of the negatives of generative AI?

What are its limitations?

- LLMs are highly trained text-predictors.
- Their responses are based on probable language.
- Factual information may therefore contain inaccuracies that sound plausible but are often entirely incorrect.
- For example, LLMs may invent quotes, references, or coding libraries.
- LLMs only have access to the data they were trained on and therefore don’t have access to current information.
- LLMs cannot perform complex computations and are not perfect at interpreting language.
- They simply predict the most likely sequence of words.


Start
Does it matter if the output is true?
NO
YES
Safe to use ChatGPT
Do you have expertse to verify that the output is accurate?
YES
NO
Are you able and willing to take full responsibility (legal, moral, etc.) for missed inacuracies?
YES
NO
Possible to use ChatGPT*
Unsafe to use ChatGPT
*but be sure to verify each output word and sentence for accuracy and common sense

Does it matter if the output is inaccurate?

What are the broader ethical issues?

Regulation

There is currently no regulation of generative AI. Its rapid development has caused apprehension for many leading figures and calls to pause its development.

Privacy and data

Generative AI is built on information scraped from the internet without permission and often in violation of intellectual property rights. OpenAI has no data regulation policy and may collect sensitive information from users through prompts.

What are the broader ethical issues? Contd..

Cognitive bias
Generative AI has no ethical principles. It is a tool that collects information from databases and texts it processes on the internet. Its predictions repeat the cognitive biases found in that information.

Environment
Training LLMs produce significant carbon emissions. Researchers suggest that training ChatGPT-3, for example, generated 552 tons of carbon dioxide. Equivalent to 123 petrol cars driven for one year.

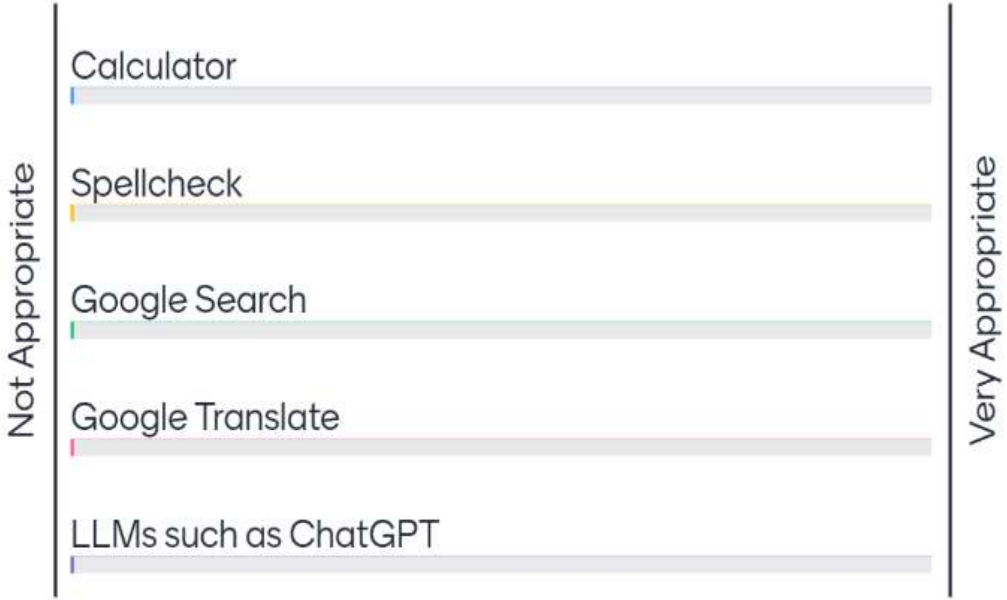
Human rights
Generative AI requires invisible human labour to build and cleanse. OpenAI employed workers in Kenya in gruelling conditions for less than $2 an hour.

1. Is it appropriate to use generative AI?

Option 1: Discuss student perspectives on different tools

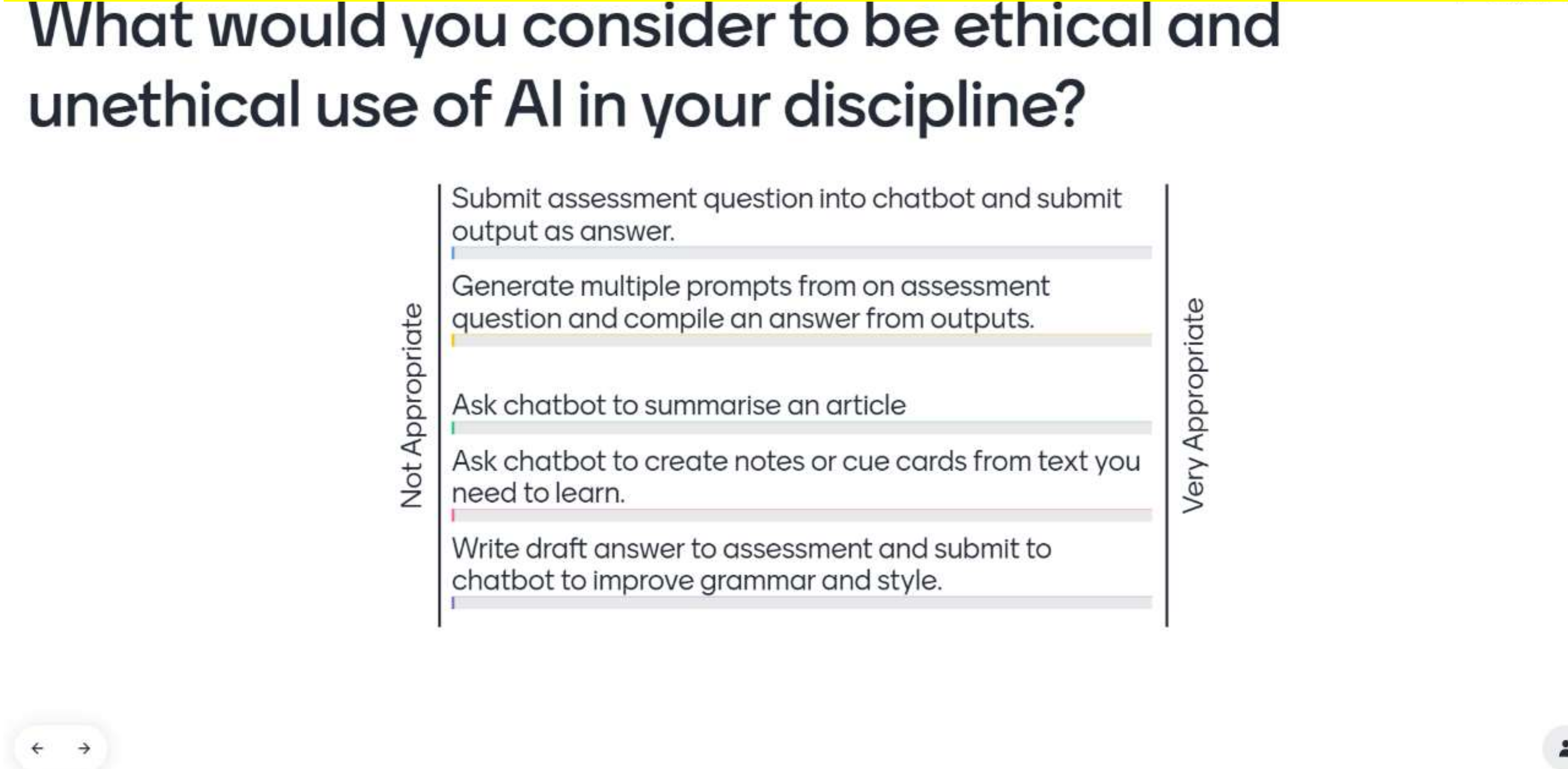
Is it appropriate?

Is it appropriate to use these tools in your discipline?



Option 2: Discuss student perspectives on ethical usage of generative AI

Option 2: Discuss ethical usage of AI



What would you consider to be ethical and unethical use of AI in your discipline?

Submit assessment question into chatbot and submit output as answer.

Generate multiple prompts from on assessment question and compile an answer from outputs.

Ask chatbot to summarise an article

Ask chatbot to create notes or cue cards from text you need to learn.

Write draft answer to assessment and submit to chatbot to improve grammar and style.

Option 3: Critique of an AI-generated output in your discipline

Can ChatGPT succeed in an exam in Univ?

- Here is an answer that was generated on ChatGPT-4 based on the prompt: Is it helpful to view nations as real or imagined?
- Spend five minutes reading the answer in full. What are your first impressions? Is it an engaging answer? Does it answer the question? Are there any mistakes or “hallucinations,” and do you feel confident that you could spot them?
- Now turn to our undergraduate marking criteria. Where might this answer meet the criteria, and at what level? Where does it fail to meet the criteria?

AI Generated Answer

Is it helpful to view nations as real or imagined?

The question of whether nations are real or imagined entities constitutes a central debate in the field of nationalism studies. This discussion grapples with the intricate interplay between tangible elements such as territory, economy, and governance, and the intangible aspects, such as shared culture identity, and historical memory that shape a nation. Three influential theorists – Benedict Anderson, Ernest Gellner, and Anthony D. Smith – provide distinct perspectives on this debate.

Anderson, in his seminal work, “Imagined Communities,” introduces the concept of nations as socially constructed ‘imagined’ entities, bound by a shared sense of identity despite their citizens rarely meeting face-to-face. Gellner, on the other hand, situates nations in the context of industrialization and modernization, suggesting that they are ‘real’ entities that emerge to meet the homogenization demands of modern societies. Meanwhile, Smith bridges these views with his ethno-symbolist approach, arguing that nations are both ‘real’ – rooted in historical ethnic communities – and ‘imagined’ – constructed through shared symbols and collective memory.

AI Generated Answer contd…

In this essay, we will delve into each theorist’s perspective to better understand whether it is more beneficial to view nations as ‘real’ tangible entities or ‘imagined’ social constructs, and what these perspectives can offer in our understanding of national identities. The exploration of these three viewpoints will illuminate the multifaceted nature of nations, underlining the complexity inherent in answering whether nations are real or imagined.

Benedict Anderson’s groundbreaking concept of “imagined communities” provides a compelling framework for understanding nations primarily as socially constructed entities. In his words, “the nation is imagined because the members of even the smallest nations will never know most of their fellow-members, meet them, or even hear of them, yet in the minds of each lives the image of their communion” (Anderson, 1983, p. 6).

Can ChatGPT succeed in an exam ? (contd.)

Argument and Writing	Evidence and Knowledge	Concepts and Historiography
Analytical introduction that sets the parameters and defines key terms.	Comprehensive knowledge and sophisticated understanding of subject	Good command of concepts relevant to the question
Logical structure enables close engagement with question throughout	Primary and secondary sources provide concise historical detail in ways which are concise, relevant, and analytical	Ability to analytically deploy concepts within argument
Develops coherent and cogent argument that culminates in a clear conclusion.	Moves between generalisations and detailed discussions with confidence and originality	Extensive reading and critical analysis of the historiography
Fluent, coherent, grammatical, accurate writing with appropriate references and bibliography	Independent use of knowledge and evidence that goes beyond paragraphing of other historians	Evaluates the type of historical argument found in the literature, including an appreciation of methodology
Excellent command of scholarly vocabulary to sustain complex argument	Critically reflects on nature and scope of evidence available	Identifies contradictions and disagreements within historiography and positions argument within debate

Can ChatGPT solve physics problems?

- Consider the following 2-part physics problem, what do you think of ChatGPT’s answer to the first part?

(1) If a deployment of stratospheric aerosol geoengineering adds 10 million tons of Sulphur to the stratosphere every year, and 10 million tons of sulphur is rained out per year, what is the average amount of sulphur deposited per unit area of the earth per year?

(2) If the global average rainfall rate is 3 mm/day, how much sulphur would be added to each litre of rainfall?

Can ChatGPT solve physics problems?

- **Part 1 ChatGPT Answer**

- To calculate the average amount of sulfur deposited per unit area of the Earth, we need to determine the net amount of sulfur added to the stratosphere.
- Since 10 million tons of sulfur are added each year, and 10 million tons are rained out per year, we can assume that the net addition is zero.
- Therefore, on average, there would be no additional sulfur deposition per unit area of the Earth due to this process.

Can ChatGPT4 solve physics problems?....

- **Part 2 ChatGPT Answer**

1. **Determine the total amount of rainfall per yr:**

- Average rainfall rate = 3 mm/day
- There are 365 days in a year.
- Total rainfall per year = (3 mm/day) * (365 days/year)
- = 1,095 mm/year

2. **Convert the total rainfall from millimeters to meters:**

- 1 meter = 1,000 millimeters
- Total rainfall per year = 1,095 mm/year / 1,000
- = 1.095 meters/year

3. **Calculate the volume of rainfall per year:**

- The surface area of the Earth is approximately 510 million square kilometers.
- Total volume of rainfall per year = (1.095 meters/year) * (510 million square kilometers)

4. **Convert the volume of rainfall from cubic meters to liters:**

- 1 cubic meter = 1,000 liters
- Total volume of rainfall per year = (1.095 meters/year) * (510 million square kilometers) * (1 billion liters/cubic meter)
- = 558.45 billion liters/year

5. Determine the amount of sulfur added to each liter of rainfall:

- Total sulfur added per year = 10 million tons
- Amount of sulfur added to each liter of rainfall = (10 million tons) / (558.45 billion liters/year)

2. What are University rules on academic integrity and AI?

AI and Assessments

There could be three categories to provide guidance for when and how students can use generative AI in their assessments. Each category describes a general approach with examples. You are free to adapt these categories, offer additional clarification, and include different examples. The three categories are:

Category 1: Students are not allowed to use generative AI for their assessment beyond what is specified in the rules of the Institute/University

Category 2: Students are permitted to use generative AI tools for specific purposes to assist with their assessment.

Category 3: Generative AI is an integral part of the assessment and students are supported and encouraged to use it extensively.

Departments and/or module leaders will need to decide which category to employ for their assessments in advance. This should be communicated to students in a standardised manner. Assessment cover sheets could include a statement for students to declare "I have read, understood and abided by the restrictions on the use of generative AI for this assignment."

Category 1

Students are not allowed to use generative AI for their assessment.

This category is for assessments demonstrating foundation level skills such as independently developing critical thinking skills, and applying knowledge or demonstrating fundamental skills that will be required throughout the programme.

You are not permitted to use AI tools for your assessment beyond the areas permitted by the Institute norms or in specified exceptions for students with a Statement of Reasonable Adjustment (SoRA).

The Instiute permits students to use software to check areas of academic writing such as structure, grammar, spelling, punctuation, and language translation. However, this may be considered Academic Misconduct if substantive changes to content have been made by the software, or in cases of language translation if the student is being assessed on their ability to translate or use a language other than English.

Category 2

Students are permitted to use generative AI tools for specific purposes to assist with their assessment.

- AI tools can be utilised to enhance and support the development of specific skills in specific ways, as specified by the tutor and required by the assessment.
- For instance, students might use AI for tasks such as data analysis, pattern recognition, or generating insights.
- Here the tutor should support and guide the students in the use of AI to ensure equity of experience, but the use of AI is not in itself a learning outcome.
- There will be some aspects of the assessment where the use of AI is inappropriate.

Category 3

Generative AI is an integral part of the assessment and students are supported and encouraged to use it extensively.

- AI can be used as a primary tool throughout the assessment process. Students will demonstrate their ability to use AI tools effectively and critically and critically to tackle complex problems, make informed judgments, and generate creative solutions.
- The assessment will provide an opportunity to demonstrate effective and responsible use of AI.
- The tutor should support and guide the students in the use of AI to ensure equity of experience

Acknowledging use of generative AI

Generative AI should not be included as a citation or reference (unless specified by your programme).
UCL, like most publishers, has stipulated that large language models cannot be considered an author. This is because they cannot take responsibility for their work and they do not generate original ideas. Rather, they reformulate ideas found elsewhere, without acknowledgement. It is also impossible for the reader to refer to the original source.

Instead, your use of generative AI should be acknowledged
UCL's rules on academic integrity state that you must be transparent in your use of AI. You must therefore acknowledge when you have used generative AI to assist you in the creation of your academic work, such as editing your prose, translating words or generating an image.

As a minimum, your acknowledgement should include the name, version, and URL of the generative AI tool and a brief description of the context in which the tool was used. For example:

I acknowledge the use of artificial intelligence tools in the production of this report. ChatGPT3.5 (https://chat.openai.com/) was used to provide an initial summary of focus group transcripts.

3. What are our next steps?

How might students use AI?

Role	Description	Example of implementation
Possibility engine	AI generates alternative ways of expressing an idea	Students write queries in ChatGPT and use the Regenerate response function to examine alternative responses.
Socratic opponent	AI acts as an opponent to develop and argument	Students enter prompts into ChatGPT following the structure of a conversation or debate. Teachers can ask students to use ChatGPT to prepare for discussions.
Collaboration coach	AI helps groups to research and solve problems together	Working in groups, students use ChatGPT to find out information to complete tasks and assignments.
Personal tutor	AI tutors each student and gives immediate feedback on progress	ChatGPT provides personalized feedback to students based on information provided by students or teachers (e.g., test scores).

How might students use AI? Contd..

Role	Description	Example of implementation
Study buddy	AI helps the student reflect on learning material	Students explain their current level of understanding to ChatGPT and ask for ways to help them study the material. ChatGPT could also be used to help students prepare for other tasks (e.g., job interviews).
Motivator	AI offers games and challenges to extend learning	Teachers or students ask ChatGPT for ideas about how to extend students’ learning after providing a summary of the current level of knowledge (e.g., quizzes, exercises).
Dynamic assessor	AI provides educators with a profile of each student’s current knowledge	Students interact with ChatGPT in a tutorial-type dialogue and then ask ChatGPT to produce a summary of their current state of knowledge to share with their teacher/for assessment.

Thank You

Artificial Intelligence and Ethics

Dr BK Murthy

Lecture 3

Introduction

- Ever since the birth of computation with Alan Turing, humans have put high hopes on the power of computers and artificial intelligence (AI).
- The term Artificial Intelligence is Coined by John McCarthy during the Summer workshop at Dartmouth college in 1955
- AI is expected to bring **significant** and diverse benefits to society – from greater efficiency and productivity to tackling a number of difficult global problems, such as climate change, poverty, disease, and conflict.

- AI technologies shape our societies.
- They have an enormous impact on our daily lives.
- At the same time, multiple legal and societal issues have revealed the potential of these technologies to produce undesirable impacts.
- Algorithms can enhance already existing biases.
- They can discriminate.
- They can threaten our security, manipulate us and have lethal consequences.
- People need to explore the ethical, social and legal aspects of AI systems.
- There is a common call for the ethics of AI – meaning how are we to develop and use this technology in an ethically acceptable and sustainable way?

- What are the ethical and moral principles we should adopt and follow?
- In this course, we'll take a look at the ethical issues related to contemporary AI, open up their background in philosophy and give them an interpretation in terms of computer and other sciences.
- The goal of course is to develop skills for ethical thinking.
- The course provides a guide – or a roadmap – on the ethically sustainable design, implementation and use of AI.
- It will introduce you to basic ethical concepts, their theoretical background, and their role in discussion on contemporary AI.

AI/ML

- **Artificial intelligence** is an overall term describing a set of different kinds of techniques to make computers behave in some kind of intelligent fashion.
- There is no agreed definition of AI, but in general the ability to perform tasks without supervision and to learn so as to improve performance are key parts of AI.
- **Machine learning** is a big topic in AI.
- Machine learning is a set of algorithms which by themselves learn to make decisions or to structure data.
- Supervised and unsupervised learning are based on data, while reinforcement learning is where the algorithm uses trial and error to learn to make sequences of decisions.

What is Ethics

- Ethics seeks to answer questions like
  - “what is good or bad”,
  - “what is right or what is wrong”, or
  - “what is justice, well-being or equality”.
- As a discipline, ethics involves systematizing, defending, and recommending concepts of right and wrong conduct by using conceptual analysis, thought experiments, and argumentation.

Subfields of ethics

- **Meta-ethics** studies the meaning of ethical concepts, the existence of ethical entities (ontology) and the possibility of ethical knowledge (epistemology).
- **Normative ethics** concerns the practical means of determining a moral (or ethically correct) course of action.
- **Applied ethics** concerns what a moral agent (defined as someone who can judge what is right and wrong and be held accountable) is obligated or permitted to do in a specific situation or a particular domain of action.

AI Ethics

- AI ethics is a subfield of applied ethics.
- Nowadays, AI ethics is considered part of the ethics of technology specific to robots and other artificially intelligent entities.
- It concerns the questions of how developers, manufacturers, authorities and operators should behave in order to minimize the ethical risks that can arise from AI in society, either from design, inappropriate application, or intentional misuse of the technology.

Main Concerns of Ethics of Technology

- Immediate, here-and-now questions about, for instance, security, privacy or transparency in AI systems
- Medium-term concerns about, for instance, the impact of AI on the military use, medical care, or justice and educational systems
- Longer-term concerns about the fundamental ethical goals of developing and implementing AI in society

From Machine ethics to the ethics of AI

- AI ethics was taken to mean mostly machine and roboethics.
- These cover the study of the ethical codes of artificial moral agents.
- As research fields, they are based on a scenario where machines can one day be responsible for ethically relevant choices, and can even be possibly considered as ethical agents or autonomous moral agents.
- As a comparison, animals are generally not considered moral agents.
- We don’t judge a squirrel’s behaviour as right or wrong, and we don’t assume they have the capacity to know the difference.

From Machine ethics to the ethics of AI…..

- Machine and roboethics span from the development of ethically responsive autonomous vehicles to the design of ethical codes for moral autonomous agents.
- Isaac Asimov,1942 famously proposed “three laws of robotics” that would guide the moral action of machines:
  - A robot may not injure a human being or, through inaction, allow a human being to come to harm.
  - A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
  - A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

- These days, AI ethics is a more general field, and closer to engineering ethics:
- we don’t have to assume the machine is an ethical agent to analyze its ethics.
- Research in the field of AI ethics ranges from reflections on how ethical or moral principles can be implemented in autonomous machines to the empirical analysis on how trolley problems are solved, the systematic analysis of ethical principles such as fairness, and the critical evaluation of ethical frameworks.

Values and Norms

- Values and norms are the basic elements of ethics.
- The concept “value” means, roughly, the degree of importance of a thing or an action.
- Values provide ideals and standards with which to evaluate things, choices, actions, and events.
- In ethics, the focus is primarily on moral values, although other types of values – economic, aesthetic, epistemic (or knowledge-related) – are sometimes relevant morally
- For example, economic factors may play morally significant role, if economic decisions have morally significant consequences to people.

Intrinsic and extrinsic values

- Values can be divided into **extrinsic** (also called “instrumental”) and **intrinsic** values.
- For example, money has extrinsic or instrumental value. Money is valuable only because one can use it for other things, such as to provide better medical care for the people.
- These things, in turn, may be good for what they lead to: for example, for better health.
- And those things, in turn, may be good only for what they lead to – for example the better quality of life.
- Intrinsically valuable things are typically “big moral values” – happiness, freedom, wellbeing.
- These are things that are good as they are.
- For some, they also explain the “goodness to be found in all the other things”.

Norms

- Norms are value-based principles, commands and imperatives – such as the sets of AI guidelines.
- They tell what one should do, or what is expected of someone.
- Norms may be prescriptive (encouraging positive behavior; for example, “be fair”) or
- Proscriptive behavior (discouraging negative behavior; for example, “do not discriminate”).

Type of Norms

- Some norms are merely statistical regularities:
- one notices that many computer scientists tend to wear black T-shirts.
- Regulations imposed by the authorities
- Some norms are social norms;
- They tell what people in a group believe to be appropriate action in that group.
- Regulations imposed by the authorities so that your actions should not cause inconvenience to others

- Moral norms are prescriptive or proscriptive rules with obligatory force beyond that of social or statistical expectations.
- For example, “Do not use AI for behaviour manipulation” is a moral norm.
- Norms may also be legal norms.
- Importantly, a legal norm may not be a moral norm, and vice versa.
- Simply, the fact that “X is a law” does not make it a moral principle.
- Instead, one can always ask: “Is this law morally acceptable or not?”

Hume´s guillotine: Facts, value and norms

- Normative claims do not describe how the world is.
- Instead, **they prescribe how the world should be**.
- This is, they imply “ought-to” evaluations, in distinction to sentences that provide "is" types of assertions.
- For instance, a sentence "This machine-learning system is a black-box system" is descriptive,
- while a sentence “Machine-learning systems should be transparent" is normative.

- Importantly, facts do not dictate our norms.
- As Scottish philosopher, David Hume (1711–76) states it, one should not make normative claims about what should be, based only on descriptive statements about what is.
- This does not mean that facts do not take any part in our moral consideration, but that you cannot get from an “is” to an “ought” without the use of some purely normative value statement along the way.
- This principle is known as “Hume´s guillotine”.

Hume´s guillotine Principle

- Hume´s guillotine principle states that moral norms or claims cannot be justified only by appealing to facts.
- As Hume remarks, one cannot derive the “ought from is”.
- For example, the fact that “there is a biased data set” does not alone imply that the data should (or shouldn’t) be biased.
- Instead, moral attitudes depend on other ethical considerations and preferences, not just mere facts.

- Why are we concerned with the issue of biased data?
- Well, the problem clearly is not the fact that there are biased data.
- The real problem is that biases may enhance discrimination.
- Importantly, Hume´s guillotine does not claim that facts don’t matter. They do.
- The point is that facts (alone) don´t solve ethical problems.
- Instead, ethical problems require genuinely ethical discussion, too.

A Framework for AI Ethics

- Traditionally, technology development has typically revolved around the functionality, usability, efficiency and reliability of technologies.
- However, AI technology needs a broader discussion on its societal acceptability.
- It impacts on moral (and political) considerations.
- It shapes individuals, societies and their environments in a way that has ethical implications.

- The interpretation of ethically relevant concepts can change with technologies (consider what “privacy” meant before social media).
- Furthermore, when new technologies are introduced, users often apply them for purposes other than those originally intended.
- This reforms the ethical landscape, and forces us to reflect and analyze the ethical basis of technology over and over again.

Ethical Frameworks

- Ethical frameworks are attempts to build consensus around values and norms that can be adopted by a community – whether that’s a group of individuals, citizens, governments, businesses within the data sector or other stakeholders.
- Various organisations have participated in developing an ethical framework for AI.
- Naturally, their views differ in some respects, but there’s also been an emerging consensus to them.
- According to a recent study (Jobin et al 2019), AI ethics has quite rapidly converged on a set of five principles.

Principles of AI Ethics

- non-maleficence
- responsibility or accountability
- transparency and explainability
- justice and fairness
- respect for various human rights, such as privacy and security

- Should we use AI for good and not for causing harm? (the principle of beneficence/ non-maleficence)
- Who should be blamed when AI causes harm? (the principle of accountability)
- Should we understand what, and why AI does whatever it does? (the principle of transparency)
- Should AI be fair or non-discriminative? (the principle of fairness)
- Should AI respect and promote human rights? (the principle of respecting basic human rights)

- The rest of this course will focus on these principles of AI ethics.
- We will analyze what these concepts imply and how they can be interpreted, in the fashion of traditional philosophy: concept analysis.
- We will also look at how these concepts are being applied in practice, discuss their problems and mention some open questions regarding these principles.
- We will look at the project of AI ethics as a whole. We will be asking the “cui bono” question: who is AI ethics for, and who or what is left out?

- Lastly, we want to note that when speaking of AI and the social implications, AI ethics is the first on the list.
- But there are other theoretical frames for looking at ethical codes for algorithmic, data-driven systems.
- For example, questions of the social implications of AI come up in fields like algorithmic cultures, gender studies and media studies, amongst numerous others.
- Correspondingly, the cognitive and psychological aspects of human-machine interaction shapes the question of appropriate ethical framework for AI.
- Simply, there is a lot more to AI ethics than just data or algorithm ethics.

Thank You

AI and Ethics

Lecture 4

What should be done

- **The principle of beneficence**
- *“AI inevitably becomes entangled in the ethical and political dimensions of vocations and practices in which it is embedded. AI Ethics is effectively a microcosm of the political and ethical challenges faced in society.” --Brent Mittelstadt*
- The principle of **beneficence** says “do good”,
- The principle of **non-maleficence** states “do no harm”.
- Although these two principles may look similar, they represent distinct principles.

- Beneficence encourages the creation of beneficial AI (“AI should be developed for the common good and the benefit of humanity”),
- while non-maleficence concerns the negative consequences and risks of AI.

- AI ethics have been primarily concerned with the principle of non-maleficence.
- Discussion has focused mostly on questions of how developers, manufacturers, authorities, or other stakeholders should minimize the ethical risks such as
  - – discrimination, privacy protection, physical and social harms – that can arise from AI applications.
- Often, these discussions are stated in terms of intentional misuse, malicious hacking, technical measures, or risk-management strategies.

- Critics claim that the emphasis on non-maleficence makes ethics a matter of finding technical solutions for technical problems.
- Moral problems are seen as things that can be solved by technical “fixes”, or by good design alone.
- The wider ethical and societal context in which technical systems are embedded is forgotten.
- Many significant issues that direct the control, governance and societal dimensions of AI are ignored.

- Technology researcher Evgeny Morozov calls this “tech solutionism” – the conviction that problems caused by technology can always be fixed by more technology.
- As a result, deep and difficult ethical problems are oversimplified and unanswered.
- One of the questions is the problem of the “common good”.
- What, exactly, does that mean?
- How AI Can be useful for common good?

The common good – calculating consequences

- Suppose you are the Chief Digital Officer in Health Dept.
- You are asked to consider whether the city’s health care organisation should move from “reactive” healthcare to “preventive” healthcare.
- You read a report. It tells about novel, sophisticated machine learning systems that would help health authorities to forecast the possible health risks of citizens.

- These methods produce predictions by combining and analyzing various sources of medical and health care systems.
- By analyzing a large number of criteria data, high-risk individuals could be identified and prioritized.
- These high-risk individuals could proactively be invited to a doctor’s appointment to get proper treatment.

The benefits

- The report mentions many advantages.
- For example, sickness prevention has a lot of potential to improve the health and quality of life for citizens.
- It would allow better impact estimation and planning of basic healthcare services.
- Preventive healthcare also has the potential to significantly reduce social and healthcare costs.
- These savings, the report emphasizes, could be used for the common good.

The potential problems

- The report also includes some concerns.
- For eg, the systems raise a number of legal and ethical issues regarding privacy, security, and the use of data.
- The report asks, for example, where is the border between acceptable prevention and non-acceptable intrusion?
- Does the Health Dept./Govt have a right to use private, sensitive medical data for identifying high-risk patients?
- How is consent to be given, and what will happen to people who don’t give their consent?
- What about those people who do not give consent because they are not able to?

- The report also raises the fundamental question of the government’s role:
- if the Govt. has information about a potential health risk and does not act upon the data, is the city guilty of negligence?
- Are citizens treated equally in the physical and digital worlds?
- If a person passes out in real life, we call an ambulance without having explicit permission to do so.
- In the digital world, privacy concerns may prevent us from contacting citizens.

- What do you think about the above example?
- As a Chief Digital Officer, would you promote the use of preventive methods?
- If your answer is something like “yes, the Dept./Govt should seek an ethically and legally acceptable way to use those methods –
- there are so many advantages compared to the possible risks”, you were probably using a form of moral reasoning called "utilitarianism".

Utilitarianism

- **Utilitarianism** is a family of ethical theories.
- It conceives “benefits” as actions that maximize well-being across all affected individuals.
- Utilitarianism is a version of consequentialism, which states that the consequences of any action are the only standards of right and wrong.
- According to utilitarianists, morally right actions are the ones that produce the greatest balance of benefits over harm for everyone affected.

- Unlike other, more individualistic forms of consequentialism (such as egoism) or unevenly weighted consequentialism (such as prioritarianism), utilitarianism considers the interests of all humans equally.
- However, utilitarianists disagree on many specific questions, such as whether actions should be chosen based on their likely results (act utilitarianism), or
- whether agents should conform to rules that maximize utility (rule utilitarianism).
- There is also disagreement as to whether total (total utilitarianism), average (average utilitarianism) or minimum utility should be maximized.

- For utilitarianists, utility – or benefit – is defined in terms of well-being or happiness.
- For instance, Jeremy Bentham, the father of utilitarianism, characterized utility as "that property… (that) tends to produce benefit, advantage, pleasure, good, or happiness…(or)
- to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered."
- Utilitarianism offers a relatively simple method for deciding, whether an action is morally right or not.

Steps to be Taken

- Firstly, we identify the various actions that we could perform
- Secondly, we estimate the benefits and harm that would result from each action
- Thirdly, we choose the action that provides the greatest benefits after the costs have been taken into account

- Utilitarianism provides many interesting ideas and concepts.
- For example, the principle of “diminishing marginal utility” is useful for many purposes.
- According to this principle, the utility of an item decreases as the supply of units increases and vice versa.
- For eg., when you start to work out, at first you benefit greatly and your results get dramatically better.
- But the longer you continue working out, each individual training session has a smaller impact.
- If you work out too often, the utility diminishes and you’ll start to suffer from the symptoms of overtraining.

- Eg.2, if you eat one sweets, you’ll get a lot of pleasure.
- But if you eat too much sweets, you may gain weight and increase your risk to all kinds of sicknesses.
- This paradox of benefits should always be remembered when we evaluate the consequences of actions.
- What is the common good now may not be the common good in the future.

The problems of utilitarianism

- Utilitarianism is not a perfect account on moral decision making.
- It has been criticized on many grounds.
- For eg, utilitarian calculation requires that we assign values to the benefits and harm resulting from our actions and compare them with the consequences that might result from other actions.
- But it’s often difficult, to measure and compare the values of all relevant benefits and costs in advance.

- "Risk" is commonly used to mean a likelihood of a danger or a hazard that arises unpredictably, or in a more technical sense, the probability of some resulting degree of harm.
- In AI ethics, harm and risks are taken to arise from design, inappropriate application, or intentional misuse of technology.
- Typical examples are risks such as discrimination, violation of privacy, security issues, cyberwarfare, or malicious hacking.
- In practice, it is difficult to compare the risks and benefits

Reasons for Not able to Compare Risks and Benefits

- **R**isks and benefits are influenced by value commitments, subjective and diverse preferences, practical circumstances, and personal and cultural factors.
- **H**arm and benefits are not static.
- The marginal utility of an item diminishes in a way that can be difficult to foresee. Moreover, a specific harm or a specific benefit may have different utility value in different circumstances.
- For eg, whether or not the faster car will be more beneficial depends on the intended use of it –
    - if it is intended to be a school bus, then we should prioritize safety, but if it is used as a racing car, then the answer may be different.

- **R**eal-world situations are typically so complex that it is difficult to foresee or compare all the risks and benefits in advance.
- For eg, let’s analyze the possible consequences of military robotics.
- Although contemporary military robots are largely remotely operated or semi-autonomous, over time they are likely to become fully autonomous.
- According to some estimates, robots reduce civilian and military casualties.
- But according to other estimates, they do not reduce the risk to civilians.
- Statistically, in the first decades of war in the 21st century, robotic weaponry has been involved in numerous killings of both soldiers and noncombatants.
- The possibility to use various techniques – such as adversarial patches (which interrupt a machine’s ability to properly classify images) – to fool and manipulate automated weapons complicates the situation by increasing the specific risk of causing harm to civilians.
- The overall level of risks is also dependent on the ease in which wars might be declared if robots are taking most of the physical risk.

- **U**tilitarianism fails to take into account other moral aspects.
- It is easy to imagine situations where developed technology would produce great benefits for societies, but its use would still raise important ethical questions.
- For eg., let’s think about the case of a preventive healthcare system.
- The system may indeed be beneficial for many, but it still forces us to ask whether fundamental human rights, such as privacy, matter. Or
- What happens to the citizen’s right not to know about possible health problems?
- Many of us would want to know if we are in a high-risk group, but what if someone does not want to know? Can a Govt. force that knowledge on them? Or,
- How can we make it sure that everyone has equal access to the possible benefits of a preventative system?

Nozick’s Utility Monster

- One of the biggest difficulties with utilitarianism is the question of utility: what is it really?
- Technically, utility is only a measure (a numeric quantity) that describes some kind of underlying “good” which we want to maximize.
- Say, pleasure, or well-being (which hedonist philosophers would claim to be the same thing).
- Pleasure is at least to some extent a subjective experience, and utility, as a measure, should transform it into an intersubjectively comparable number.
- That is a high bar to reach.

- Assuming such a measure as utility does in fact exist, philosopher Robert Nozick presents the following puzzle.
- There is a creature called the Utility Monster. Their hedonistic mind is wired so that, given any resource, they will receive more pleasure from it than any other individual would.
- They simply enjoy apples, cars, coffee, freedom, etc., more than anybody else does.
- This means that they gain more utility from them, and if we are morally obligated to maximize the utility produced by the resources we have, the conclusion is clear: everything we have to the Utility Monster. Nothing to anybody else.
- Does this make utilitarianism unpalatable?
- Is there a way for the utilitarian to argue that the puzzle Nozick posed is not really a problem?

Common good and well-being

- Despite the problems outlined earlier, the principles of utilitarianism may help us to consider the immediate and the less immediate consequences of our actions.
- One should remember that in real life, defining “common good” requires a diversity of viewpoints.

What is "well-being"?

- Often, the term “common good” is taken to be synonymous with “well-being”. But what is well-being?
- The roots of well-being research are in ancient Greece, where philosophers such as Aristotle focused on how to achieve “the good life”.
- Since then, the search for the good life has been a constant topic handled by different disciplines.
- Today, research on fields such as in psychology, economics, and social sciences addresses well-being in terms of "the biological, personal, relational, institutional, cultural, and global dimensions of life".
- These dimensions cover factors such as physical and mental vitality, social satisfaction, and a sense of personal achievement and fulfillment.

Theories of Well-being

The subjective theories.

- These focus on questions such as
  - how people feel as they go about their daily lives, or
  - how a person evaluates their lives.
- This type of psychological well-being is often described as
  - the experience of high life satisfaction,
  - high levels of pleasant emotions and
  - moods, and low levels of negative emotions and moods.

The eudaimonic theories.

- These consider well-being primarily as the outcome of positive goal pursuits.
- The eudaimonic perspective differentiates well-being from the satisfaction of desire.
- Well-being and subjective happiness should not be equated because the pleasure-producing outcomes that underlie subjective happiness do not necessarily promote wellness and well-being.
- Instead, well-being can be taken to require components such as autonomy, environmental mastery, personal growth, positive relations with others, a sense of having a purpose in life, and self-acceptance.
- These dimensions describe well-being as an overall positive evaluation of oneself, acceptance of one’s past life and individual talents as a member of a community, the belief that one’s life is meaningful, and a sense of self-determination.

The social theories.

- In these, well-being is approached in terms of social factors, such as
    - integration, contributions to social life, social coherence, and social acceptance.
- Well-being is dependent on the degree to which an individual is functioning well in their social environments.

- Surveys like The World Happiness Report provide examples of this holistic approach to well-being.
- The report is an annual publication of the United Nations Sustainable Development Solutions Network.
- It contains articles and rankings of national happiness based on respondent ratings of their own lives, which the report also correlates with various life factors.
- (As of March 2020, Finland was ranked the happiest country in the world three times in a row.)

- Moreover, researchers develop constantly novel ways to approach well-being.
- For eg, big data is nowadays utilized for well-being research in many ways.
- Contemporary methods include the more advanced analysis of demographic and socio-economic data,
- For eg. utilization of text mining tools in any written documents – such as Twitter feeds, Facebook posts, or other social media data, as well as the analysis of digital footprints and even facial features.

Coomon Good Approach for AI

- The common good approach requires that everyone should have access to the benefits of AI.
- This highlights the importance of ensuring that potential benefits of AI do not accumulate unequally, and are made accessible to as many people as possible.
- AI should be aligned with values, goals, and norms, respecting cultural and individual diversity to a sufficient degree.

- The common good is not a singular, but a plural.
- Identifying social and moral norms of the specific community in which an AI will be deployed is, thus, obligatory.
- It is the only way to bring AI’s potentially significant and diverse benefits to society and facilitate, among other things, greater well-being and welfare for all.

Thank you

Artificial Intelligence and Ethics

Lecture 5

Who should be Blamed

- What does accountability actually mean, and
- How does it apply to AI ethics?
- What does moral agency and responsibility mean and the difficulty of assigning blame
- Whom to be blamed?
- Who could be sued?
- Who could be rewarded?

Algorithms and accountability

Case Study

- In the city of Amsterdam, parking control is partially automated and in use across 150,000 street parking spaces in the city.
- Three Step Process is followed:

1. Scan cars equipped with cameras drive through the city and use object recognition software to scan and identify the license plates of surrounding cars.
2. After identification, the license plate number is checked against the National Parking Register to validate if the car has permission to park at a given location. Whenever no payment has been made for current parking, the case is sent to a human inspector for further processing.
3. A parking inspector uses the scanned images to remotely assess whether there is a special situation such as loading or unloading, or stationary cars in front of a traffic light. The parking inspector may also verify the situation on-site by scooter. Whenever there is no valid reason for non-paid parking, a parking ticket is issued.

- Parking control services provide an example of how algorithms are increasingly used for automating public services.
- As algorithms are exact, fast and precise, they often promote better service efficiency, reliability and consistency.
- Paradoxically, algorithms can also make systematic errors, be biased and cause serious harms.
- For eg., scanning systems may malfunction, or suffer from bugs.
- They may make mistakes and suggest the tickets be issued on invalid grounds.
- In these cases, who should take the responsibility – and on what grounds?

- Although we say things like "yes, it was the algorithm’s fault and
- it is responsible for the wrong decision", we do not literally mean that contemporary algorithms would be morally guilty.
- Instead, the algorithms are causal factors that underlie the decisions.
- Mere causes differ from morally responsible actions.

- Even though algorithms themselves cannot be held accountable as they are not moral or legal agents, the organizations designing and deploying algorithms can be taken to be morally responsible through governance structures.
- Thus, in the case of the city of Amsterdam, it is the human inspector that makes the final decision – and also takes responsibility.
- However, one day the human inspector may be replaced by algorithms, too. Who, then, will take responsibility?

Automated vs. autonomous decision making

Automated systems

- Typically run within a well-defined set of parameters and are very restricted in what tasks they can perform.
- The decisions made or actions taken by an automated system are based on predefined heuristics or rules.

An autonomous system

- Learns and adapts to dynamic environments, and evolves as the environment around it changes.
- The data it learns and adapts to may be outside what was considered when the system was deployed.
- Automation or autonomisation is about degree, and hence, they are continuums rather than simple yes/no situations.
- eg., a system can be said to be autonomous with respect to human control to a certain degree.

What is accountability?

- Accountability means the state of being responsible or answerable for a system, its behavior, and its potential impacts.
- Accountability is an acknowledgement of responsibility for actions, decisions, and products
- Responsibility can be legal or moral (ethical).
- **Legally**, an actor is responsible for an event when a legal system is liable to penalise that actor for that event.
- **Morally**, an actor is responsible for an act, if they can be blamed for the action.
- Moral and legal responsibility are different things.
- They do not always coincide; an agent can be legally responsible even if they were not morally responsible, and vice versa.
- Here we´ll focus only on moral aspects of responsibility.

Dimensions of Accountability

- The question of determining the responsibility –
  - which individuals (or groups) are accountable for the impact of algorithms or AI?
  - Who is responsible for what effect within the overall socio-technical system?
- A feature of the societal system that develops, produces, and uses AI
- A feature of the AI system itself

Who should be blamed – and for what?

- In ethics, accountability is closely related to the concept of “moral agency”.
- A moral agent is “an agent who is capable of acting with reference to right and wrong."
- Importantly, only moral agents are morally responsible for their actions.

Actions and omissions

- Philosophically, a moral agent is primarily responsible for their own actions (“acts”).
- Sometimes agents are also responsible for not-doings, “omissions”.
- So, if I kill someone, I am responsible for that act.
- If I just let someone die, I am responsible for not-helping (omission), even if I did not actively kill someone.

- Omissions and actions are not morally equal.
- It is morally less bad to omit a thing than to perform an act:
  - It is worse to actively kill someone than to let them die.
  - But this doesn’t make omissions morally right.
  - However, we cannot be responsible for all of the things we do not do.
  - Instead, we are responsible for only those things which we’ve deliberately and knowingly chosen to do or omit.

Autonomy

- Philosophically, moral responsibility requires
    - moral autonomy and
    - the ability to evaluate the consequences of actions.
- “Moral autonomy” means the agent’s capacity to impose the moral code on oneself in a self-governed way.

Further, autonomy requires:

- The capacity to rule oneself without manipulation by others and the ability to act, without external or internal constraints
- The authenticity of the desires (values, emotions, etc) that move someone to act

Sufficient cognitive skills –

- meaning an agent must be able to evaluate, to predict and
- to compare consequences of their actions and,
- also, to estimate motives that drive action by using ethically meaningful criteria

Moral responsibility

- Immanuel Kant is one of the most famous European moral philosophers.
- For Immanuel Kant, practical reason — our ability to use reasons to choose our own actions — presupposes that we are free.
- Actions are based on our own will to utilize a moral law to guide our decisions.
- For Kant, and Kantians, the claim is that this capacity (to impose the moral law upon ourselves) is the ultimate source of all moral value.

- So, according to Kant, we owe ourselves moral respect in virtue of our autonomy.
- But we owe similar respect to all other persons in virtue of their capacity.
- Hence, we are obliged to act out of fundamental respect for other persons in virtue of their autonomy.
- In this way, autonomy serves as both a model of practical reason in the determination of moral obligation and as the feature of other persons who deserve moral respect from us

The problem of individuating responsibilities

- Accountability is often taken as a legal and ethical obligation on an individual or organisation to accept responsibility for the use of AI systems, and to disclose the results in a transparent manner.
- This formulation presupposes a “power-relation”. It individuates who is in control and who is to be blamed.
- However, it has turned out to be notoriously difficult to set specific criteria on how, exactly, the responsibilities should be individuated, directed and defined.
- In many countries, there are on-going debates on these questions.
- International actors, such as European Union and G7 have addressed them as open challenges.

Why is it so difficult to set criteria on who is responsible?

- **1.** The quality of responsibilities differ.
- An actor is responsible for a specific action or omission,
- but the quality of responsibility is dependent on the stakeholder.
- Thus, although by choosing an action you may commit the responsibility,
- the quality of responsibility is dependent also on your properties.
- Intelligent technologies complicate this more.

- As we delegate more and more decision making tasks and functions to algorithms, we also shape decision making structures.
- AI is augmenting our intelligence by giving us more computational power, allowing better predictions and enhancing our sensory apparatus.
- Human and machines become cognitive hybrids.
- They cooperate cognitively (thinking) and epistemically (knowledge), both at the individual and collective level.
- This creates systemic properties.

- It is often thought that it is sufficient that a human stays “in-the-loop” or “on-the-loop” –
    - meaning at some point of decision making, a human individual would be able to monitor or intervene in the artificial system.
- However, as algorithms enter into decision making, say, in public sector governance, the collective decision making can take a very complex and highly distributed form.
- To individuate and address the factors in a way that would guarantee that a human stays in/on-the-loop may be really difficult.

- **Secondly**, technology can also take the persuasive form: it influences and controls people.
- A classical example is the beeping sound of seat belts.
- In many cars, if the seat belts are not fastened, it will cause a constant beeping sound.
- This can be taken as a form of controlling influence—in this case a kind of coercion.
- The driver can only stop the sound by fastening the belt. Contemporary algorithmic applications can have more and more such features; they propose, suggest and limit the options.

- But, an action is done voluntarily only if the action is done intentionally (the one acting is “in control”) and is free from controlling influences.
- Is the driver free from controlling influences, if the seat belt system forces him to react to the beeping sound?
- Or, are we free from control, if the algorithms decide whose pictures we´ll see in the dating sites, or what music we are about to listen to?
- What, exactly, is the difference between algorithmic suggestion, control, or manipulation?

- Naturally, persuasive technology should comply with the requirement of voluntariness to guarantee autonomy.
- Algorithms complicate this issue, since the voluntariness presupposes a sufficient understanding of the use of specific technology.
- But, what does it mean to “understand”, and what is the sufficient degree, really?
- What is the correct reading of “understandability” – “transparency”, “explainability” or “auditability”?
- How much, and what, exactly, a user should understand about the technology?
- When can one genuinely estimate, whether or not they want to use that particular technology?

Thank You

AI and Ethics

Lecture 6 Transparency in AI

Transparency in AI

- Why is transparency in AI important and
- what major issues are affected by transparency – and
- what are some of the risks associated with transparency in AI systems?
- How the Transperency help for Social Good?

Principle of Transparency

- Imagine a facial recognition system called MYFACE.
- MYFACE is used for security purposes in the airport.
- Usually it works perfectly, but one day it starts to miscategorize individuals as potentially dangerous.
- As a result, several innocent people are arrested.
- Would it be important to know why the system made all these mistakes?
- Should we be able to explain why it made mistakes? And
- why would this matter?

Transparent AI

- Some contemporary machine learning systems are so-called “black box” systems, meaning we can’t really see how they work.
- This “opacity”, or lack of visibility, can be a problem
  - if we use these systems to make decisions that have an effect on individuals.
- Individuals have a right to know how critical decisions –
  - such as who gets accepted for a loan application,
  - who gets paroled, and who gets hired – are made.
- This has led many to call for “more transparent AI”.

Transparency in AI

- Transparency is a property of a system that makes it possible to get certain information regarding a system’s inner workings.
- But what information that is, and whether it is ethically relevant, depends largely on the ethical issue we are trying to answer.
- Transparency itself is ethically neutral and is not an ethical concept. Instead, it constitutes an ideal.
- Transparency is something that can manifest in many different ways, and something that can present a solution for underlying ethical questions.
- In this sense, transparency is relevant at least to the three issues:

The justification of decisions.

- Good governance in public or private sectors involves non-arbitrariness of decisions.
- This is applied to any kind of decision-making that has an ethically or legally relevant effect on individuals.
- Non-arbitrariness means access to justifications about “why was this decision reached, and on what grounds?”
- Furthermore, the capacity to contest and appeal are crucial.
- This represents a demand to right wrongs.
- This also leads to accountability

A right to know

- According to human rights, people are entitled to have explanations on how decisions were made so that
- they can maintain genuine agency, freedom and privacy
- Freedom entails the right to get answers to questions such as “How am I being tracked?
- What kind of inferences are being made about me?
- And how, exactly, have the inferences about me been made?”

A moral obligation to understand the consequences of our actions

- As a community, we also have a responsibility for managing risks.
- There is a moral obligation, up to some reasonable level, to understand and predict the consequences of the kinds of technologies one brings into the world.
- That is, saying “we can’t understand now what it will do” is not a valid argument for unleashing a system that causes harm.
- Instead, it is our moral duty to explore the possible risks.

- These three points can all be summarized as calls for sufficient information.
    - Justification for the Decisions/Actions
    - A right to know
    - A moral obligation to understand the consequences of our actions
- Do we know whether and to what extent this algorithmic decision is justified?
- Do I know how inferences about me are made?
- To what extent I am responsible for the actions of the system, and
- how much I should know about the inner workings of the system to be able to take that responsibility?

What is transparency?

- Transparency can be defined in multiple ways.
- There are a number of neighboring concepts that are sometimes used as synonyms for transparency –
- including “explainability” (AI research in this area is known as “XAI”),
- “interpretability”, “understandability”, and “black box”.

Transparency

- Transparency is, roughly, a property of an application.
- It is about how much it is possible to understand about a system’s inner workings “in theory”.
- It can also mean the way of providing explanations of algorithmic models and
- decisions that are comprehensible for the user.
- This deals with the public perception and understanding of how AI works.
- Transparency can also be taken as a broader socio-technical and normative ideal of “openness”.

Transparency

- There are many open questions regarding what constitutes transparency or explainability, and
- what level of transparency is sufficient for different stakeholders.
- Depending on the specific situation, the precise meaning of “transparency” may vary.
- It is an open scientific question, whether there are several different kinds, or types, of transparency.
- Transparency can refer to different things whether the purpose is to, say, analyze the legal significance of unjust biases or
- to discuss them in terms of features of machine learning systems.

Transparency as a property of a system

- As a property of a system, transparency addresses how a model works or functions internally.
- Transparency is further divided into
- “simulatability”
  - (an understanding of the functioning of the model),
- “decomposability”
  - (understanding of the individual components), and
- algorithmic transparency
  - (visibility of the algorithms).

What makes AI system a “black box”?

Complexity.

- In contemporary AI-systems, operation of a neural network is encoded in thousands, or even millions, of numerical coefficients.
- Typically the system learns their values at the training phase.
- Because the operation of the neural network depends on the complicated interactions between these values,
- it is practically impossible to understand how the network works even if all the parameters are known.

Difficulty of developing explainable solutions.

- Even if the used AI models support some level of explainability, additional development is required to build explainability to the system.
- It may be difficult to create a user experience for careful yet easily understandable explanations for the users.

Risk concerns.

- Many AI algorithms can be fooled if an attacker carefully designs an input that causes the system to malfunction.
- In a highly transparent system, it may be easier to game the system to come up with strange or unwanted results.
- Thus, sometimes systems are intentionally designed as black boxes.

- Given that many of the most efficient, current deep learning models are black box models,
- researchers seem to assume it is highly unlikely that we would be able to develop them as fully transparent.
- Because of this, the discussion focuses on finding the “sufficient level of transparency”.
- Would it suffice if algorithms offered people a disclosure of how algorithms came to their decision and provide the smallest change “that can be made to obtain a desirable outcome” ?
- Eg., if an algorithm refuses someone a social benefit, it should tell the person the reason, and also what he or she can do to reverse the decision.

- The explanation should tell, for instance, what the maximum amount of salary to be approved is (input), and
- how decreasing the amount will impact the decisions made (manipulation of the input).
- But the problem is that the right to know also applies to situations where the system makes mistakes.
- Then, it may be necessary to perform an autopsy on the algorithm and identify those factors that caused the system to make mistakes.
- This can’t be done by only manipulating the inputs and outputs.

- This illustration depicts a very simplified AI model tasked to recognize all cats in data consisting of all kinds of animals.
- The model has inferred two patterns that make up a cat.
- To the model, they're just numbers, but to us, they look like describable patterns.
- However, patterns and features inferred can look quite complicated to us.
- Moreover, transparency serves many other functions in contemporary debates on machine learning models.
- It can be relevant for developing legislation or for ensuring public trust in AI.
- To handle these issues the notion of transparency in AI is typically given a wider definition in terms of “comprehensibility”.

Transparency as comprehensibility

- The comprehensibility – or understandability – of an algorithm requires that
- one should explain how a decision was made by an AI model in a way that is sufficiently understandable to those affected by the model.
- One should have a concrete sense of how or why a particular decision has been arrived at based on inputs.
- However, it is difficult to translate algorithmically derived concepts into human-understandable concepts.
- In some countries, legislators have discussed whether public authorities should publish the algorithms they use in automated decision-making in terms of programming codes.

- However, most people do not know how to make sense of programming codes.
- It is thus hard to see how transparency is increased by publishing codes.
- Would it be more helpful to publish the exact algorithms?
- In most cases, publishing the exact algorithms does not bring a lot of transparency either,
- especially if you do not have the access to the data used in a model.

- Nowadays, cognitive and computer scientists develop human-interpretable descriptions of how applications behave, and why.
- Approaches include, the development of data visualization tools, interactive interfaces, verbal explanations or meta-level descriptions of the features of models.
- These tools can be extremely helpful for making AI applications more accessible.
- The fact that comprehensibility is based on subject and culture-dependent components complicates this more.
- Eg. the logic of how visualizations are interpreted – or
- how the inferences are made on them – varies across cultures.
- Thus, tech developers should pay attention to the sufficient understanding of the visual language they use.

- Moreover, much is dependent on the degree of data or algorithmic literacy, or the knowledge of contemporary technologies.
- In some cultures, the vocabulary of contemporary technology is more familiar, but in many others they may be completely novel.
- To increase the understandability, there is clearly a need for significant educational efforts in improving algorithmic literacy – for example on “computational thinking”.
- This user literacy will have a direct effect on transparency in terms of the ordinary users’ basic understanding of AI systems.
- It may actually provide the most efficient and practical way to make the boxes less black for many people.

How to make models more transparent?

- The black box problem of artificial intelligence is not new.
- Providing transparency for machine learning models is an active area of research.
- Roughly speaking, there are five main approaches:

Use simpler models.

  - This, however, often sacrifices accuracy for explainability.

Combine simpler and more sophisticated models.

  - While the sophisticated model allows the system to do more complex computations, the simpler model can be used to provide transparency.

Modify inputs to track relevant dependencies between inputs and outputs.

  - If a manipulation of inputs changes overall model results, these inputs may play a role in the classification.

Design the models for the user.

- This requires using cognitively and psychologically efficient methods and tools for visualizing the model states or directing attention.
- Eg, in computer vision, states in intermediate layers of the models can be visualized as features (like heads, arms, and legs) to provide a comprehensible description for image classification.
- Researchers have also developed methods for directing “attention” towards the parts of the input that matter the most.
- These can be visualized to highlight the parts of an image or a text (so-called “weights”) that contribute the most to a particular recommendation.
- **Follow the latest research**. A lot of research is ongoing on various aspects of explainable AI – including the socio-cognitive dimensions – and new techniques are being developed

Transparency and the risks of openness

- Transparency often denotes a modern, ethico-socio-legal “ideal” a normative demand for the acceptable use of technology in our societies.
- It is a reflection of the ideal of “openness”, that is framed in terms of “open government”, “open data”, “open source/code/access”, as well as “open science” .
- Transparency considerations are needed to mitigate the equal distribution of scientific advancements
- so that the benefits of AI development can be made accessible for all people.

- Paradoxically, the ideal of openness can lean to harmful consequences, too.
- Eg., the transparency of social media platforms has led to several instances of misuse and democratic challenges.
- Transparency can create security risks.
- Too much transparency may lead to leaking of privacy-sensitive data into the wrong hands.
- Or the more that is revealed about the algorithms and the data, the more harm a malicious actor can cause.
- Algorithms can be hacked, and information may make AI more vulnerable to intentional attacks.
- Entire algorithms can also be stolen based simply on their explanations alone.

Thank You

AI and Ethics

Lecture 7 AI Respect and Human Rights

Human Rights and AI

- What are human rights, and how do they tie into the current ethical guidelines and principles of AI?
- The three rights of particular importance to AI:
  - the right to privacy,
  - security, and
  - inclusion

Human Rights

- During the COVID-19 pandemic, governments have struggled to find effective policy-making strategies for exiting lockdown in a safe way.
- According to epidemiologists, opening up society requires efficient tracking, tracing, and monitoring.
- In many cases, this has led to the utilization of various tracing and tracking apps.
- These apps have raised several concerns about privacy and security.
- Critics have seen them as the first steps towards the algorithmic surveillance of citizens

- In London, authorities decided to try something new.
- Together with scientists, they developed methods for “capturing activity over London” to better understand the city’s level of activity.
- In a project called Odysseus, authorities get information about the distribution of activities in London by combining machine learning algorithms, statistical time-series analysis, and image processing.
- This information about the activity in the streets of London can be utilized for the safe reopening of streets and for public health planning.

- In Odysseus, the data comes from a wide range of sources.
- Odysseus combines aggregated, anonymized mobile phone data, anonymized credit card transactions, satellite navigation data, and data from sensors and traffic cameras in the streets.
- This data is used to create counts of vehicles, cyclists and pedestrians, and to indicate the density and impacts of social distancing.
- Special attention is paid to the anonymization of data so that individuals cannot be identified.

- The right for a safe environment is one of these.
- Odysseus provides an example of how AI can be used in a way that respects and promotes the right to safety or to a healthy environment.
- At the same time, the project must take other rights – such as the right for privacy – into account.
- In London these concerns were taken seriously.
- To secure privacy, Odysseus is designed in a way that all the data is anonymized and individuals cannot be identified from the images taken by the traffic cameras.

- Privacy and security have raised a lot of media attention.
- They are important, but it’s necessary to consider the impact of AI on the full spectrum of fundamental human rights and freedoms, too.
- How will AI impact on the right to education and work, or for a fair trial, to fair and open elections, to freedom of speech, and to assembly and demonstration?
- And what about special groups, such as children?
- But first, let’s discuss what human rights are.

What are human rights?

- Human rights form the foundation of the current ethical guidelines and principles of AI.
- This makes human rights a fundamental component of contemporary AI ethics.
- As rights, human rights are **universal**: all humans are entitled to have them.
- One does not have to be a particular kind of person or a member of some specific community to have human rights.
- Human rights are **norms** that protect all people, everywhere from political, legal, and social abuses.

- **Civil and political rights**, such as the right to life, liberty, and property, freedom of expression, pursuit of happiness, and equality before the law
- **Social, cultural and economic rights**, including the right to participate in science and culture, the right to work, and the right to education
- The role of human rights is to protect people's ability to form, construe, and pursue their own conceptions of a worthwhile life – it's not just about the ability to live “in liberty, happiness and well-being”.

What is a human right?

- A human right is a norm which can exist on different levels:
- a shared norm of actual human moralities
- a justified moral norm supported by strong reasons
- a legal right at the national level (where it might be referred to as a “civil” or “constitutional” right)
- a legal right within international law

Universal Declaration of Human Rights

- UDHR is a document which was drafted by representatives with different legal and cultural backgrounds from all regions of the world.
- The declaration was proclaimed by the United Nations General Assembly in Paris on 10 Dec. 1948 as a common standard of achievements for all peoples and all nations.
- Conceptually, human rights are grounded in agency and autonomy.
- They have an ethical priority: if they compete with other considerations such as economic wealth, national stability or some other factor, human rights should be prioritized.

- In the context of AI, this prioritization implies the following requirements:

- AI applications that could clearly violate human rights should not be used
- AI applications that prevent people from enjoying their human rights or actively put them at risk of human rights violations should not be used

- However, human rights have certain context-sensitive properties that allow individuals to prioritize a specific human right if needed.
- Some rights are more fundamental than the others. For example, when the right to life conflicts with the right to privacy, the right to privacy will generally be outweighed.
- In recent years, privacy and security concerns have dominated the discussion on AI and human rights.
- Emerging combinations of big data analytics, surveillance technologies and developing biometric recognition methods have recently received significant media and policy attention.
- Also, the right to equality and inclusion has raised a lot of public discussion. In the next section, we’ll take a brief look at these discussions.

Examples of human rights: privacy, security, and inclusion

- **Privacy**
- Privacy concerns are raised, for example, by digital records which contain information that can be used to infer sensitive attributes (age, gender or sexual orientation), preferences, or religious and political views.
- Biometric data also raises privacy concerns, as it can reveal details of physical and mental health.
- Often the real worry is not the data itself, but the way the data can be used to manipulate, affect, or harm a person.
- Ethically, privacy is related to personal autonomy and integrity. Following the principles set out by John Locke, a right to control our own personal lives has been seen as central to our autonomy.
- If that right is taken away, it violates something fundamental about our psychological and moral integrity.

- Many have proposed the principle that people should have control over their own data – and that data concerning them should not be allowed to be used to harm or discriminate against them.
- According to some, this right to have “full control over one’s own data” should be a human right.
- But what, exactly, is your “own data”?
- Is it the raw data, or the collected and analyzed data?
- If the data is used for secondary purposes, is it still your data? Or, as Wachter and Mittelstadt (2019) remark,
- does the content of inferences that can be drawn from your data belong to your “own data”?

- Wachter and Mittelstadt (2019) propose that the right for the control of your own data should be reformulated as a right for the “right to reasonable inferences”.
- According to them, it is crucial that we can also control the “high-risk inferences” that can be made about us through big data analytics.
- These inferences are privacy-invasive or reputation-damaging, or have low verifiability (in the sense of being predictive or opinion-based) while being used for important decisions.

General Data Protection Regulation (GDPR)

- The General Data Protection Regulation (GDPR) is a legal framework.
- It sets guidelines for the collection and processing of personal data from individuals who live in the European Union.
- The GDPR's aim is to give individuals control over their personal data.
- Any information that relates to an individual who can be directly or indirectly identified is “personal data”.
- This includes names, social security numbers and email addresses. Location information, biometric data, ethnicity, gender, web cookies, and political or religious beliefs can also be personal data.
- Pseudonymous data (data that does not directly identify an individual but can be connected to them) can also fall under the definition if it’s easy to individuate someone from it.

- The data subject must give specific, unambiguous consent to process the data.
- Consents must be “freely given, specific, informed and unambiguous.”
- Data subjects can withdraw previously given consent whenever they want.
- Children under 13 can only give consent with permission from their parent.
- The GDPR recognizes several privacy rights for data subjects. Their aim is to give individuals more control over the data.

GDPR Privacy Rights

The right to be informed

- A person must be told about the use of their personal data

The right of access

- It should be explained how someone's personal data is used

The right to rectification

- A person has the right to be forgotten and the data deleted

The right to restrict processing

- A person can deny the use of their personal data

Data Protection Principles in GDPR

- **Lawfulness, fairness and transparency**: Processing must be lawful, fair, and transparent to the data subject
- **Purpose limitation**: You must process data for the legitimate purposes specified explicitly to the data subject when you collected it
- **Data minimization**: You should collect and process only as much data as absolutely necessary for the purposes specified
- **Accuracy**: You must keep personal data accurate and up to date

Data Protection Principles in GDPR….

- **Storage limitation**: You may only store personally identifying data for as long as necessary for the specified purpose
- **Integrity and confidentiality**: Processing must be done in such a way as to ensure appropriate security, integrity, and confidentiality (for example by using encryption)
- **Accountability**: The data controller is responsible for being able to demonstrate GDPR compliance with all of these principles

How to protect privacy: data anonymization methods

- The GDPR permits organisations to collect anonymized data without consent, use it for any purpose, and store it for an indefinite time – as long as organisations remove all identifiers from the data.

Generalization

- A method that deliberately removes some of the data to make it less identifiable.
- Data can be modified into a set of ranges or a broad area with appropriate boundaries.
- You can remove, for example, the street address while including the information about the town name.
- In this way, you can eliminate some of the identifiers while retaining a degree of data accuracy.

How to protect privacy: data anonymization methods

Pseudonymization

- A data management and de-identification method that replaces private identifiers – names, ID-codes – with fake identifiers or pseudonyms,
- For eg., replacing the identifier “Shanti Devi” with “Saara”.
- Pseudonymization preserves statistical accuracy and data integrity.
- The modified data can be used while still protecting data privacy.

- **Synthetic data**
- A method for using created artificial datasets instead of altering the original dataset.
- The process involves creating statistical models based on patterns found in the original dataset.
- One can use standard deviations, medians, linear regression or other statistical techniques to generate the synthetic data.

- Data-anonymization can be challenging.
- There are also methods for “de-anonymization”.
- De-anonymization methods attempt to re-identify encrypted or obscured information.
- For example, cross-reference anonymized information with other available data in order to identify a person, group, or transaction.

Safety and security

- The right to safety is a norm protecting individuals from physical, social and emotional harms, including accidents and malfunctions.
- Security means safety from malicious and intentional threats.
- As a right, safety creates a moral obligation to design our products, laws and environment in such a way that safety can be protected even in unconventional circumstances or impairments.

AI as an existential threat

- The conversation around AI as an existential threat takes a highly speculative and future-oriented stance towards artificial intelligence.
- It focuses on asking what kind of threats to humanity are posed by AI systems if they become too complex to control.
- However, the plausibility of a future of super-intelligent AI has been called into question, both by philosophers and technologists.
- As things stand, there is no reason to assume that superintelligence will emerge from developing contemporary algorithmic methods.

Safety in AI

- Safety in AI is the practical question of designing systems which behave in a safe and predictable manner.
- As AI systems are integrated into ever-widening areas of life, it becomes more important that the systems are well designed to account for the complexity of the world.
- A very practical and already existing example of this is lane guard technology, which uses machine learning to prevent cars from veering outside of their lanes.
- Machine learning researchers have found that some lane detection algorithms are quite easy to confuse with rogue road markings, causing the car to veer off the road by following the fake lane markings.

- One could argue that the right to safety obligates technology producers to account for these kinds of scenarios:
- the fact that the environment was not ideal does not excuse the system malfunctioning.
- This feature is called **robustness** – the capacity of the system to work predictably under new and unpredictable circumstances.

- The ethically – and legally – significant question is “what are the acceptable limits to robustness?”
- It is conceivable that there are a set of circumstances so incredible that even if the system’s safety cannot be assured,
- We can concede that “nobody could have realistically seen that coming”.
- Where this limit is, though, is a difficult problem, and definitely not one that is exclusive to AI or even technology.

Case: Caged pavements – AI safety and environmental uncertainty

- A difficult problem for autonomous vehicles is the complex unpredictability of the urban traffic environment.
- While AI-driven vehicles are constantly being developed to include better ways to model their surroundings, even a small group of individuals – all performing their own movement goals within a shared space – will create a constellation that is difficult to predict.
- When technical solutions in the cars are too far off, another way to approach the issue is to contain the uncertainty in the environment.

- In a New York Times column, Eric A. Taub proposed a solution: by enclosing pavements in cages, with traffic-light-synced gates at crossings,
- So that the complex traffic environment is simplified to become more understandable to autonomous vehicles and therefore safer.
- However, this safety comes at an obvious cost: limiting the freedom of pedestrians, and a redistribution of accountability.
- This means we should look at the intersecting limits of the right of safety vs freedom.
- Which one is more important?

- A further interesting line of thought that can be traced here is the criminality of what in the US is called “jaywalking”, or
- Walking across the road at locations without zebra crossings.
- The concept of jaywalking did not exist until the roads were reconceptualized with motor vehicles as the primary users.
- How comparable is this to the thought of caging pavements?

Producing safety with AI

- Can AI make the world safer?
- Can AI make the world feel safer?
- And safer for whom?
- Robotization can provide an example of this concept in practice.
- The work of handling hazardous materials or
- working in hazardous environments can be delegated to robots, protecting the health of human (or animal) workers.

- AI-powered surveillance is used in many domains: in public spaces, in law-enforcement work through predictive policing, and in domestic life through products like Amazon’s Ring.
- “Still, the social implications of being recorded have not changed: when we walk into a store, we generally expect that the presence of cameras won’t affect us.
- We expect that our movements will be recorded, and we might feel self-conscious if we notice a camera, especially if we were doing anything that we feel might attract attention.
- But unless something dramatic occurs, we generally understand that the videos in which we appear are unlikely to be scrutinized or monitored.”

- Constant surveillance produces **“chilling effects”**.
- That is, the awareness that our actions are constantly watched limits our true freedom to act in the world.
- Imagine that whenever you leave your house, you are tailed by two police officers.
- They never interact with you, just follow ten meters behind you.
- You will probably feel unsettled and unable to go about your day as you normally would.
- In this way, safety is sometimes at odds with personal freedom and privacy.

A safe and healthy environment: AI and climate change

- Safety also means the right to a safe and healthy environment.
- Nowadays, this right is threatened by climate change.
- The effects of climate change are already visible – storms, droughts, fires, and flooding have become more common, more frequent and more devastating.
- Global ecosystems are changing.
- They all impact the environment on which our existence depends.
- The report on climate change, 2018 estimated that the world will face catastrophic consequences unless global greenhouse gas (GHG) emissions are eliminated within thirty years.

- AI could be a powerful tool for tackling climate change.
- It can be used as a resource for monitoring, understanding, and predicting the consequences of climate change.
- AI can accelerate the development of more ecologically sustainable societies.
- It can be used to design green cities, environment-friendly transportation, to reduce the ecological impact of industry, and
- Design equipment that can help study and maintain the diversity of ecosystems.

- Many potential problems are associated with the deployment of AI – for instance, innovations that seek to reduce greenhouse gas emissions may actually increase energy consumption and emissions.
- Given the data and resource-intensive character of contemporary AI, the technology itself still struggles with energy consumption and carbon footprint.
- One must also pay attention to the environmental impact of raw material extraction for supporting the manufacturing of AI technologies, which can be significant.

Summary

- Safety plays into AI technologies in multiple different ways.
- These all raise questions about the balancing of normative values: while calls to make “AI for good” sound promising,
- In practice the enactment of rights and normative values in technological systems often collides with the many conflicting interests and deep injustices existing in the world.
- When evaluating safety, it is then important to evaluate what other rights intersect in practice and ask, “safety for whom?”

Thank you