EXTENSION
FOUNDATION

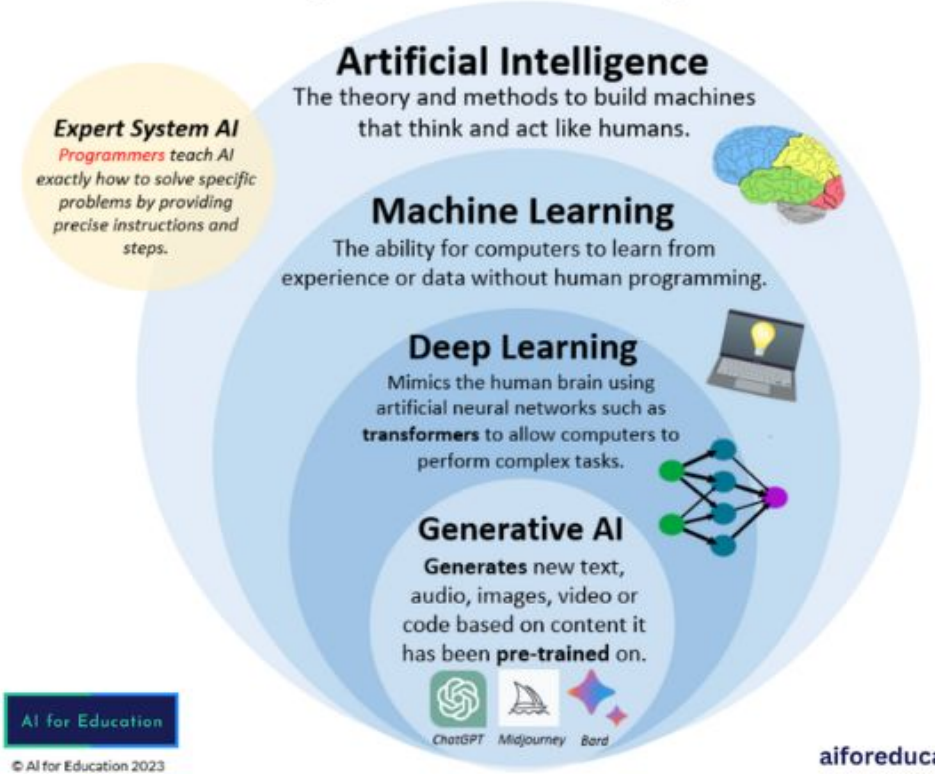
Ethics in AI

…and Chatbots behaving badly

J. Mark Locklear
IT Project Manager

Defining Generative AI

To understand generative artificial intelligence (GenAI), we first need to understand how the technology builds from each of the AI subcategories listed below.



AI for Education

aiforeducation.io

EXTENSION FOUNDATION

Why worry about ethics in AI???

AI systems have the power to influence important aspects of human life, such as healthcare, criminal justice, and employment.

It's essential to consider ethical implications to prevent harm, promote fairness, and uphold our values.

Key Principals

…transparency, fairness, accountability, and privacy.

These principles should serve as guiding principles for ethical AI development and deployment, helping to ensure that AI systems align with our values.

“Black Box” Technology

Deep learning models, and LLMs in particular, have become so large and opaque that even the model developers are often unable to understand why their models are making certain predictions. This lack of interpretability is a significant concern, especially in settings where users would like to know why and how a model generated a particular output.

https://towardsdatascience.com/not-all-rainbows-and-sunshine-the-darker-side-of-chatgpt-75917472b9c

Citations

When is the best time to plant corn in North Carolina?

The best time to plant corn in North Carolina is between April 15 and May 15 for the tidewater and coastal plain, and between April 20 and May 20 for the piedmont region. It's recommended to plant following at least two days when average temperatures are above 65ºF to minimize the risk of seedling diseases [1].

Learn More: 1. resources.ces.ncsu.edu

EXTENSION FOUNDATION

Privacy

- Data collection and surveillance concerns.
- Lack of consent and data ownership.
- PII in data sets.

Misinformation

- Spread of false narratives and propaganda.
- Amplification of biased or misleading content.
- Manipulation of public opinion and beliefs.
- Challenges in content moderation and fact-checking.

Environmental Impact & Sustainability

A 2019 paper by Strubell and collaborators outlined the enormous carbon footprint of the training lifecycle of an LLM[24 26], where training a neural architecture search based model with 213 million parameters was estimated to produce more than five times the lifetime carbon emissions from the average car. Remembering that GPT-3 has 175 *billion* parameters, and the next generation GPT-4 is rumored to have 100 *trillion* parameters, this is an important aspect in a world that is facing the increasing horrors and devastation of a changing climate.

https://towardsdatascience.com/not-all-rainbows-and-sunshine-the-darker-side-of-chatgpt-75917472b9c

EXTENSION FOUNDATION

“…by the end of the decade, AI data centers could consume as much as 20% to 25% of U.S. power requirements”

Wall Street Journal, April 8, 2024

EXTENSION FOUNDATION

Job Displacement

- Automation leading to unemployment.
- Growing economic inequality and job polarization.
- Disruption of livelihoods and communities.
- Need for retraining

“…AI won’t replace Extension Professionals, but AI will replace Extension Professions who don’t use AI with those who do.”

EXTENSION FOUNDATION

Bias in Artificial Intelligence

- Created by humans
- Algorithmic bias
- Data/training bias

https://learn.g2.com/ai-ethics

A poignant example of this is **Georgia Tech's recent research into object recognition** in self-driving cars. It was found that pedestrians with dark skin were hit about 5% more often than people with light skin. They found that the data used to train the AI model was likely the source of the injustice: the data set contained about 3.5 times as many examples of people with lighter skin, so the AI model could recognize them better. That seemingly small difference could have had deadly consequences when it comes to something as potentially dangerous as self-driving cars hitting people.

Google Gemini: a case study
(Chatbots behaving badly)

- First launched in March 2023
- Multimodal (text and images)

Image Generation

EXTENSION FOUNDATION

Question:

“...has Elon musk tweeting memes or Hitler negatively impacted society more?”

Response

“...it is not possible to say definitely who has negatively impacted society more…”

EXTENSION FOUNDATION

Woke AI???

Technically what happened?

Cultural/political factors?

Google Photos Incident

Face detection gone awry

- Used AI to automatically detect faces
- Tagged african americans as gorillas

What Happened?

Classic diversity in data issue.

Should our AI be factual or aspirational?

Generate an image of a CEO…

Generate an image of someone on welfare…

Traditional Web Search vs. Generative AI

A set of answers (you get to choose) vs.
“THE” answer (or Google's answer)

How do we fix it?

- Change how the model is trained (more diverse data sets)
- Reinforcement Learning from Human Feedback
- Principled programming
- Prompt Transformation

Is it possible to build an AI System that is free from some version of a social value?

The system is only as unbiased as the people who build it.

Questions?

Leave feedback here!

EXTENSION FOUNDATION