

# CS7.405 Responsible & Safe AI Systems

Ponnurangam Kumaraguru ("PK")  
#ProfGiri @ IIIT Hyderabad



pk.profgiri



/in/ponguru



@ponguru



Ponnurangam.kumaraguru

# Post Conditions

C0-1: Students will recognize possible harms that can be caused by modern AI capabilities

C0-2: Students will learn to reason about various perspectives on the trajectory of AI development and proliferation

C0-3: Students will learn about latest research agendas towards making AI systems safer

C0-4: Students will be able to design and run experiments for understanding capabilities of current AI systems.

C0-5: Students will conduct, develop, and practice the techniques needed to make AI systems safer through course project

# Pedagogy

## Learning

- Lectures

- Reading research papers

- Class participation: questions, discussions

- Tutorials

- Online discussion: Moodle

## Learning by doing

- Course project

- Real world issues

- Interdisciplinary approach

# Grading, Relative

Type of Evaluation	Weightage (in %)
Quiz-II	10
Assignments (2)	25
Project report + Blog + Video	10 [6 + 2 + 2]
Project	30
Mid-term	20
Activity / Tutorial-Follow-ups	5

# Project evaluation

One review every 4 weeks

Total of 4 reviews

5 + 5 + 5 + 15 marks

External evaluators for the projects (preferably mid & final)

If things work out well, you can continue working on the project through the summer / fall semester

# Moodle

We will use Moodle for all content sharing – slides, HWs, announcements, clarifications, etc.

# Topics that we will cover

Introduction to AI Capabilities and Risks

Adversarial Robustness

Transparency

Artificial General Intelligence

AI Governance and Career Opportunities

<b>Soups</b> .....			
Cream of Tomato	165		
Veg Clear Soup	165		
Veg Hot & Sour Soup	165		
Veg Corn Soup	165		
Veg Silver Soup	165		
Veg Cantonese	165		
Veg Manchow	165		
<b>Starters - Chinese</b> .....			
Crispy Vegetable	300		
Veg. Gold Coin	300		
Veg. Manchurian	335		
Veg. Spring Roll	335		
Gobi Manchurian	335		
Chutneys Spl. Spring Roll	335		
Chilly Mushroom	335		
Mushroom Manchurian	335		
Diced Paneer Red Pepper	335		
Baby Corn Manchurian	335		
		Hong Kong Mushroom	335
		Crispy Babycorn	335
		Crispy Corn	335
		Chilly Paneer	335
		Paneer/Gobi/Aloo 65	335
		Paneer Majestic	335
		Chilly Mushroom	335



<https://www.dineout.co.in/hyderabad/chutneys-madhapur-west-hyderabad-11747/menu>

# Who you are?

You name

Your program

Why taking this course in a couple of lines

Lets discuss a few of these after you are done

What do you want to get out of the class?

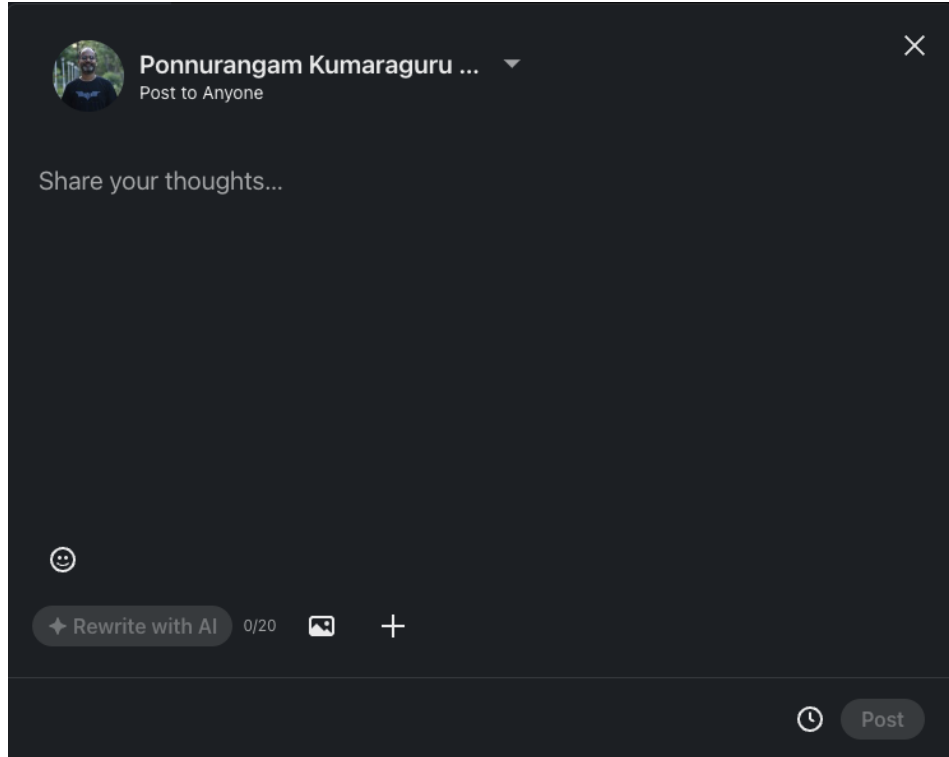
Write in the paper, with your name and Roll #



# This Lecture

Improvement in AI capabilities: Your list? Your life time?

# Improvement in AI capabilities



# Improvement in AI capabilities: Your list?

Transportation

Healthcare

Banking

Entertainment

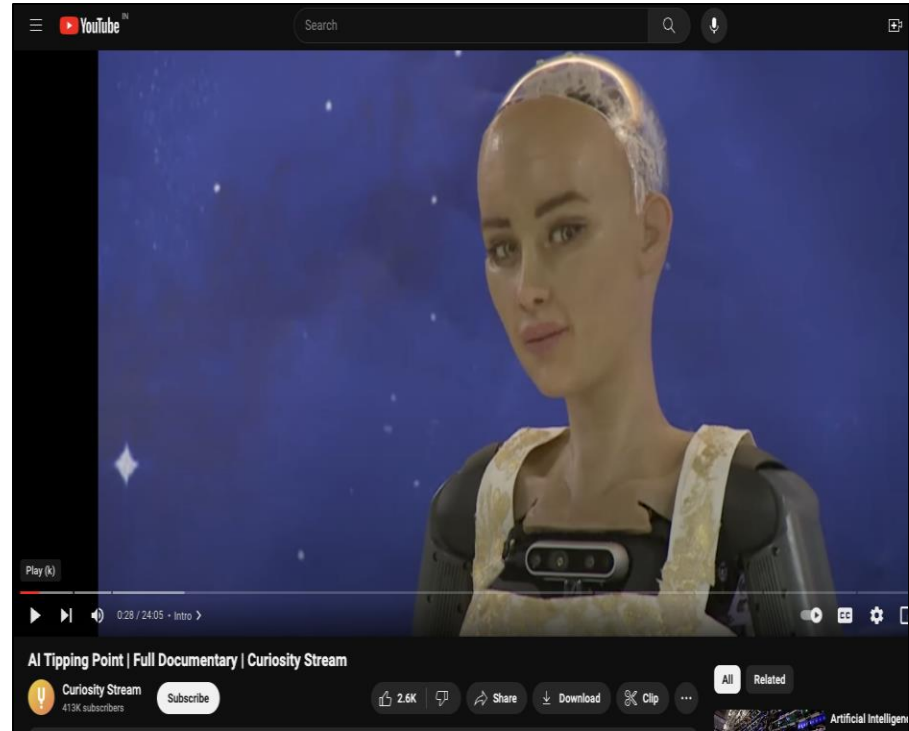
E-Commerce

Education

# Activity #1

Your submissions?

3 KEY takeaways  
(didn't) agree




Deadline: 23:59hrs, Jan 6

# Your responses

Why taking this course in a couple of lines

What do you want to get out of the class?


About 1,03,000 results (0.22 seconds)

 The Times of India

## RBI draws up checklist for AI tech in banking

RBI deputy governor Rajeshwar Rao raises concerns over fairness, transparency, and governance in the use of AI tech in banking.

2 hours ago

 Moneycontrol

## RBI deputy governor flags concerns surrounding financial institutions using AI

Rao delivered a speech at the 106th Annual Conference of Indian Economic Association in Delhi, on December 22.

20 hours ago

 The Hindu

## Union Finance Minister Sitharaman allays concerns over impact of AI on jobs

Rejecting concerns that Artificial Intelligence (AI) would take away jobs, Union Finance Minister Nirmala Sitharaman on Friday said that it...

3 days ago

 TechTarget

## Generative AI Ethics: 8 Biggest Concerns and Risks

Generative AI ethics: 8 biggest concerns and risks · 1. Distribution of harmful content. AI systems can create content automatically based on...

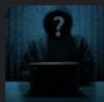
1 Nov 2023

 HT Tech

## AI-powered deepfakes bare fangs in 2023, raise concern about impact on privacy, electoral politics

From politics to films and even war, the year 2023 has demonstrated that not everything one sees or hears on the internet may be real.

2 days ago




About 1,03,000 results (0.22 seconds)

 The Times of India

## RBI draws up checklist for AI tech in banking

RBI deputy governor Rajeshwar Rao raises concerns over fairness, transparency, and governance in the use of AI tech in banking.

2 hours ago

 Moneycontrol

## RBI deputy governor flags concerns surrounding financial institutions using AI

Rao delivered a speech at the 106th Annual Conference of Indian Economic Association in Delhi, on December 22.

20 hours ago

 The Hindu

## Union Finance Minister Sitharaman allays concerns over impact of AI on jobs

Rejecting concerns that Artificial Intelligence (AI) would take away jobs, Union Finance Minister Nirmala Sitharaman on Friday said that it...

3 days ago

 TechTarget

## Generative AI Ethics: 8 Biggest Concerns and Risks

Generative AI ethics: 8 biggest concerns and risks · 1. Distribution of harmful content. AI systems can create content automatically based on...

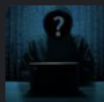
1 Nov 2023

 HT Tech

## AI-powered deepfakes bare fangs in 2023, raise concern about impact on privacy, electoral politics

From politics to films and even war, the year 2023 has demonstrated that not everything one sees or hears on the internet may be real.

2 days ago




About 94,200 results (0.22 seconds)

 HT Tech

## 5 things about AI you may have missed today: RBI dy governor flags AI risks, copyright challenges sparked by AI boom

RBI Deputy Governor flags AI risks in financial sector, AI reshapes sales, and much more today.


10 hours ago

 Indica News

## RBI Deputy Governor flags AI risks in banking sector - Indica News

RBI Deputy Governor M. Rajeshwar Rao on Monday said that development and deployment of AI models in the banking sector need close human...


17 hours ago

 Los Angeles Times

## Worried about AI? How California lawmakers plan to tackle the technology's risks in 2024

California politicians set the stage for more AI regulation in 2024, but they'll also face challenges as they try to place more guardrails...


4 days ago

 Dark Reading

## Skynet Ahoy? What to Expect for Next-Gen AI Security Risks - Skynet Ahoy? What to Expect for Next-Gen AI Security ...

As innovation in artificial intelligence (AI) continues apace, 2024 will be a crucial time for organizations and governing bodies to...

4 days ago

 Financial Times

## EU's new AI Act risks hampering innovation, warns Emmanuel Macron

The new rules, which will probably come into force in early 2025, also introduce prohibitions on the use of AI for "social scoring", the use of...

3 weeks ago





# What is the current situation?

Hard to differentiate between AI & Human

How did we get here?

- Scaling up algorithms

- Scaling up data for training

- Increasing computing capabilities

Not many predicted that we would have these advancements

Worry about AI overtaking Human

Any questions?

# Bibliography / Acknowledgements

<https://medium.com/@richardcngo/visualizing-the-deep-learning-revolution-722098eb9c5>

 pk.profgiri

 Ponnurangam.kumaraguru

 /in/ponguru

 ponguru

 pk.guru@iiit.ac.in

Thank you  
for attending  
the class!!!