

# CS7.405 Responsible & Safe AI Systems

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



pk.profgiri



/in/ponguru



@ponguru



Ponnurangam.kumaraguru

# Protocol



# Who you are?

CND

ECD

EHD

ECE

CLD

CSE

CSD

???

# Who am I?

- ~~Assistant~~ Associate Professor of Computer Science
- Ph.D. from School of Computer Science, Carnegie Mellon University (CMU)
- Research interests
  - Computational Social Science
  - Social (Societal) Computing
  - Privacy & Security in Social Media
- Courses I teach
  - Data & Applications (1), 4+
  - Online Privacy (1)
  - Privacy and Security in Online Social Media (8), 4+
  - Designing Human Centered Systems (5), 4+
  - Research methods / Advanced research methods (2), 4+
  - Foundations of Computer Security (5), 4+
  - Big Data & Policing (1), 4+



# My Philosophy

“It is our choices [project, group members, course, professor to work with 😊 etc.], Harry, that show what we truly are, far more than our abilities.”

**“It is our choices, Harry, that show what we truly are, far more than our abilities.”**

B.Tech. Orientation  
Dec 18, 2021  
IIIT Una



[https://en.wikipedia.org/wiki/Albus\\_Dumbledore](https://en.wikipedia.org/wiki/Albus_Dumbledore)

# My Philosophy

Work ~~hard~~ smart, play hard

*Brick walls are there for a reason. The brick walls are not there to keep us out. The brick walls are there to show how badly we want something.*

If the requests are reasonable, most of the course related decisions (deadline extension, ...) , will be student-friendly, so, concentrate on the content





Ponnurangam Kumaraguru is with P. J. Narayanan and 12 others.

February 14 · 🌐

Received 4.2 (on 5) on course evaluation for Privacy and Security in Online Social Media at IIIT Hyderabad, taught in Fall 2018. Parameters for evaluation - Course content, Class conduct, and Instructor. Super satisfying to see that students liked the course! 23 students took the course. Thanks to every student for your involvement and creative project work! Special thanks to Shreya, for TA-ing the course and to all project evaluators, and admin support! #IIITHStudentsRock #ProfGiri #IIITH #Sabbatical 100+ pics from the final poster presentation <https://www.facebook.com/media/set/...>



Saksham Suri, Dheeraj Reddy Pailla and 66 others



Ponnurangam Kumaraguru is with Vedant Das Swain and 31 others.

December 14, 2015 · 👥

Received 4.5/5 in "Course Instruction", 4.3 in "Course Management", and 4.4 in "Instructor to Students Feedback" for the CSE 648: Privacy and Security in Online Social Media, which I taught this Fall semester. 30 out of 38 students filled the course feedback. HUGE thanks to every student who took the course; was fun teaching this course for the first time in a different format. Thanks to my Teaching Assistants (Srishti, Prateek) and external evaluators. #LovingMyFacultyLife #IIITDStudentsRock

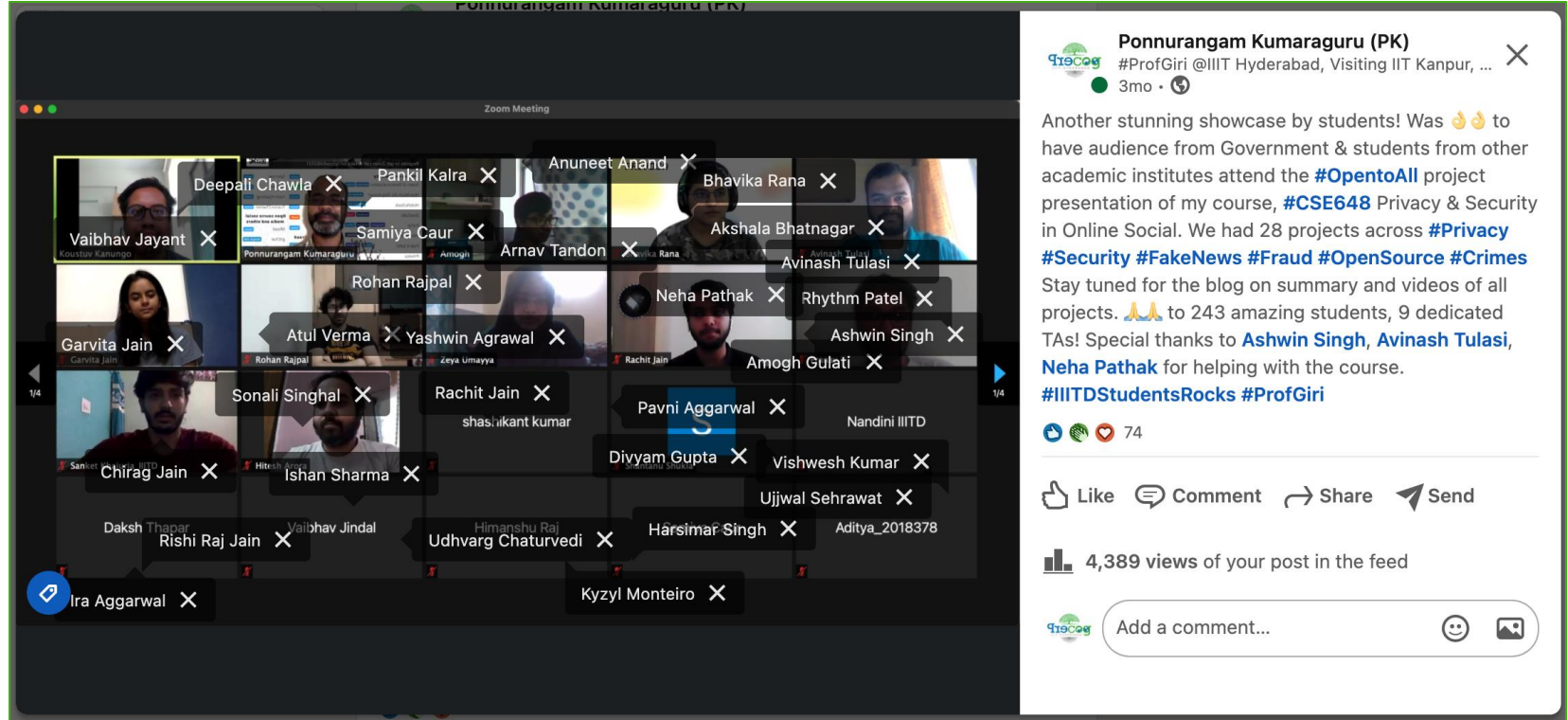


Srishti Gupta, Prateek Dewan and 122 others

3 Comments



# Fun at the end of the semester!





# Why should I teach this course?

Super thrilled and excited about the topic

Realization on the importance of systems being responsible & safe

Multiple Ph.D. students, Masters / DD, UG, RAs work with me in this area

There are plans for summer school around this topic on campus

# 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

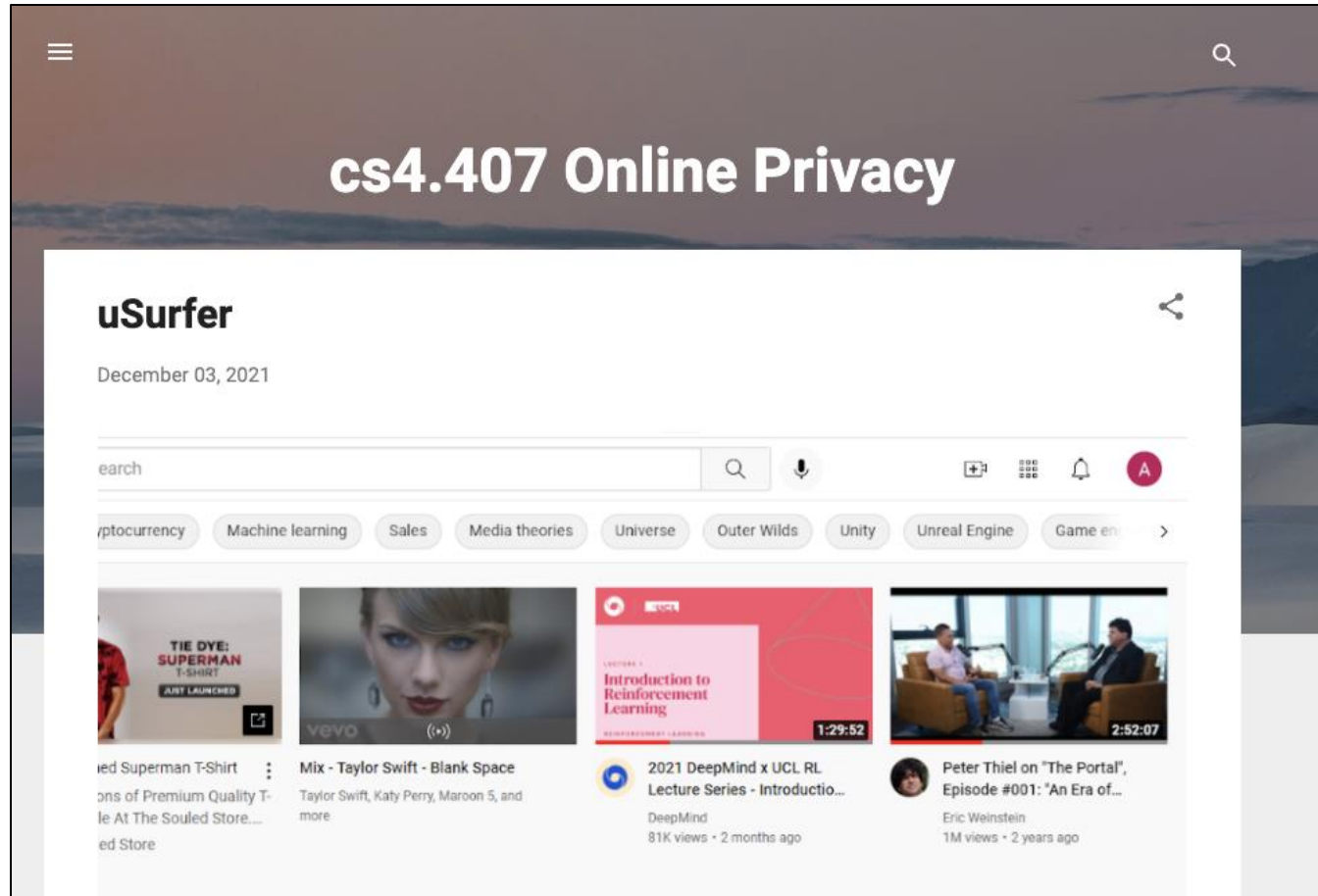
One review every 4 weeks

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



# TAs







Project Mentors (optional)

# Disclaimer

Given the nature of the topic, it is exciting and relevant, so we can look at what is happening during the semester and pick them up

- Will be happy to change (a bit) the content as we proceed

- Discuss about GAB in Fall 2018 semester

- Will be great to have these going on Moodle?

I am excited, willing to learn, and modify as it works for you

If you focus on looking at some interesting problems around, you and if we can address it through the class that will be super cool!

# Other details

## Course timings & Classroom

Mondays 0830 – 1000 hrs

Thursdays 0830 – 1000 hrs

## Course materials

No prescribed book

Papers will be circulated

All course materials will be shared on Moodle

## Office hours

We will announce it soon

PK & TAs

Will try and share slides before the class

# Guest Lectures

Wendy Hall, Univ of Southampton

Neel Nanda, Google DeepMind

Arun Jose, Independent Researcher

Prof. Ravi Balaraman, IIT Madras

Daniel Paleka, PhD Student, ETH Zurich

Dr. Adam Gleave, CEO FAR AI

Dr. Dan Hendrycks, Director of Center for AI

Dr. Ethan Perez, Research Scientist, Anthropic

Prof. Vincent Conitzer, Carnegie Mellon University

Anybody more?

Mostly evening IST!

# Plagiarism

What is it?

Copying HWs

Any content taken from another source without citation

First time caught, zero on that submission

Second time caught will be directed to academic committee

Caught in final report, one grade lower

Whatever policy from IIITH

# Topics that we will cover

Introduction to AI Capabilities and Risks

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>

# Moodle

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



# Course website


[Home](#) [Research](#) [People](#) [Publications](#) [Resources](#) [Teaching](#) [Events](#) [News](#) [Blog](#)

## Responsible and Safe AI Systems

Spring 2024

### Course Staff

**Instructor:** Ponnurangam Kumaraguru "PK"



### Course Topics

Module 1: Introduction to AI Capabilities and Risks  
Module 2: Adversarial Robustness  
Module 3: Transparency  
Module 4: Artificial General Intelligence  
Module 5: AI Governance and Career Opportunities

[Access Full Curriculum](#)


**NEW COURSE**

## Responsible & Safe AI



TOPICS TO BE COVERED



- Risks from AI Models: Toxicity, Bias, & Lies
- Artificial General Intelligence (AGI)
- Adversarial Attacks - Vision, NLP
- Representation Engineering, Model Editing & Probing
- Difficulties in Designing & Enforcing AI Regulation
- [+ More](#)

INSTRUCTED BY PONNURANGAM KUMARAGURU ("PK")  
MODE OF INSTRUCTION IN-PERSON AT IIIT HYDERABAD  
DURATION JAN - APR 2024



For any further questions or clarifications write to [pk.guru@iiit.ac.in](mailto:pk.guru@iiit.ac.in)

Also available on  





# Award

Brave penguin to take the 1<sup>st</sup> dive

Hard problem

Consistent efforts during the semester

# Service Level Agreement

Any question / clarification ask, if not urgent, will be answered in 24 hrs

If anything, urgent, feel free to attach the time in which you want the answer, we will try to respond

TAs are your 1<sup>st</sup> point of contact only on escalation, you will bring it up to me

# 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

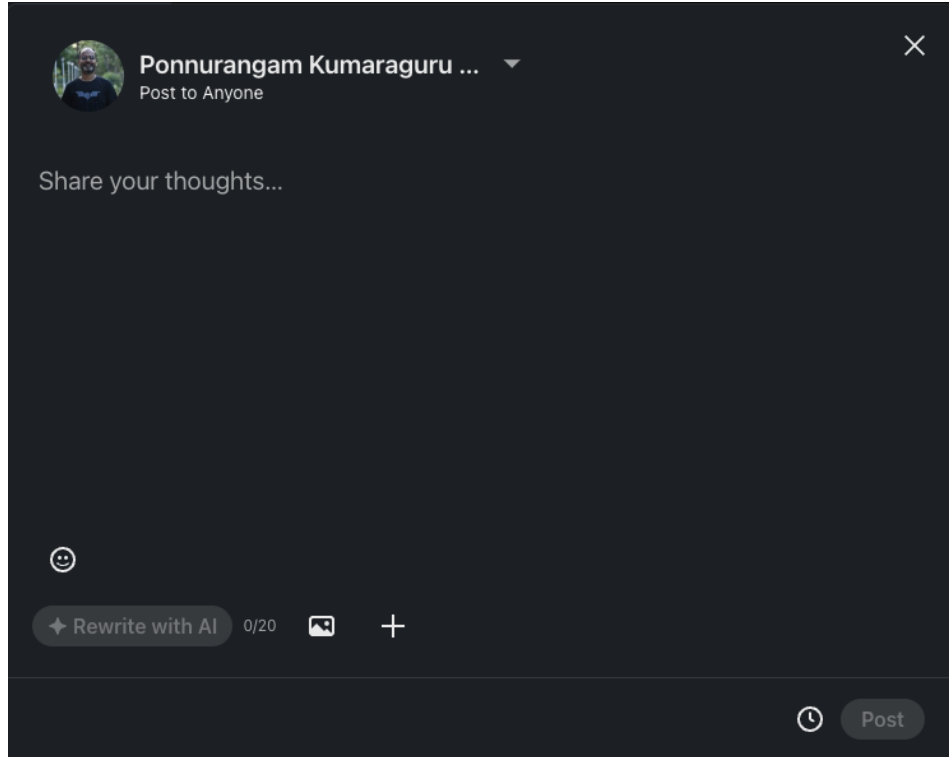
What do you want to get out of the class?

Write in the paper, with your name and Roll #

Any questions / clarifications?

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

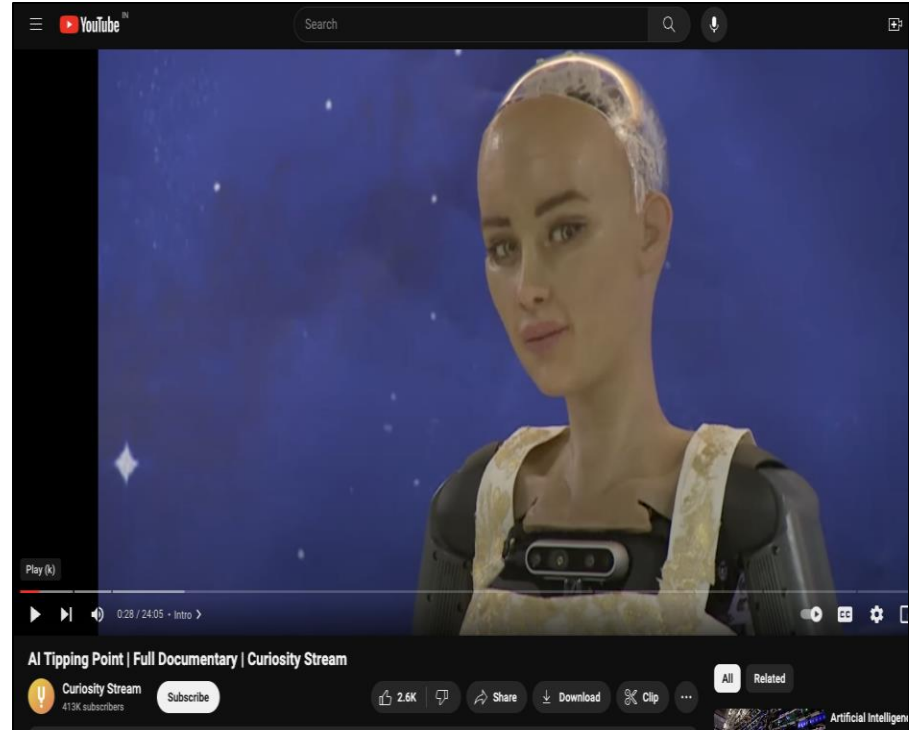
# Activity #1

AI Tipping point

<https://www.youtube.com/watch?v=1cKE12LK4Eo>

Submit 3 KEY takeaways according to you

Submit any aspect that you agree & did not agree with



Deadline: 23:59hrs, Jan 6

Any questions?

# Bibliography / Acknowledgements

 pk.profgiri

 Ponnurangam.kumaraguru

 /in/ponguru

 ponguru

 pk.guru@iiit.ac.in

Thank you  
for attending  
the class!!!