Future AI Systems Extra Credit (aka Yann LeCun Guest Lecture)

Graded

Student

Chetan Hiremath

Total Points

20 / 20 pts

Question 1

Future AI Systems

5 / 5 pts

✔ **+ 5 pts** Correct

Question 2

Your Take

5 / 5 pts

✔ **+ 5 pts** Correct

Question 3

RLHF

5 / 5 pts

✔ **+ 5 pts** Correct

Question 4

Yann's Take

5 / 5 pts

✔ **+ 5 pts** Correct

Q1 Future AI Systems

5 Points

Note: This should be done after watching the video *Objective-Driven AI: Towards AI systems that can learn, remember, reason, plan, have common sense, yet are steerable and safe*, Yann LeCun, NYU, Meta (56:47)

LINK: https://www.youtube.com/watch?v=vyqXLJsmsrk

Extra Note: While this assignment is worth Problem Set extra credit points, the material is fair game for the final -- same as any other lecture from the course.

I watched the Guest Lecture video by Yann LeCun

- (●) Yes
- ( ) No

Q2 Your Take
5 Points

Summarize in one paragraph of your own words what you learned from the video.

I have recently watched the guest lecture video that talks about objective-driven AI for AI systems that learn, remember, reason, plan, and have common sense even though they are steerable and safe. I get to learn that humans and animals learn new tasks quickly since they can learn, remember, reason, plan, and have common sense. The speaker says that machine learning is not good since the capabilities of the current learning systems are not efficient and accurate. Their behavioral skills are driven by objectives, so humans and animals can understand the overall environment of the world. Self-supervised learning is used for understanding and generating text, images, videos, 3D models, and speech. I have learned several topics like Auto-Regressive Generative Architectures, LLMs and AR-LLMs since they are used in ML. I get to know that LLMs are good at writing aids, but they have limited knowledge because they are trained from text. There are pros and cons in AI and LLMs, which are not always perfect, since they are created and trained by humans. It is good to know pros and cons of the models that are explained by the speaker. I get to know that ChatGPT thinks that it is intelligent, but it has limited knowledge and doesn't know the world and human intelligence. He talks about Objective-Driven AI that uses a model to predict the outcomes of the humans' actions. Its goal is to find a sequence that minimizes the objectives, and it is Objective-Driven AI since there is no way to break the system that is hardwired to optimize the objectives. So, it can learn models of the world and predict the proper results. Here is my summary of this video.

Q3 RLHF
5 Points

Yann mentions RLHF. What is it, and what effect did it have on the success of ChatGPT/GPTn?

Yann LeCun mentions RLHF that stands for Reinforcement Learning for Human Feedback. RLHF is a ML technique that uses human feedback to train and optimize ML models efficiently, so these models can self-learn and gather valuable data. This ML technique is used by ChatGPT/GPTn because ChatGPT collects the data of the human feedback and provides accurate and proper results when the collection of human knowledge is very large. ChatGPT is successful by RLHF because it incorporates human feedback into the reinforcement learning cycle to provide desired outcomes, so users can knows accurate results from ChatGPT.

Q4 Yann's Take
5 Points

Did Yann's model for Objective-Driven AI (or any of its elements) remind you of topics we covered in the course? How did his model/viewpoint align or not with the view of AI espoused in our course? Elaborate.

Yann LeCun's model for Objective-Driver AI reminds me of some topics like Supervised Learning, Self-Supervised Learning, and Reinforcement Learning that are covered in this course because I have recently learned these topics in this courses. Yann's viewpoint doesn't align with the view of AI that is espoused in our course because he mentions that AI models are not really great, have limited knowledge and no connection with the physical reality, don't provide factual and consistent answers, can't take recent information, and won't behave properly. This course shows that AI is a very important tool in technology, but Yann thinks that AI hallucinates and is not the smartest tool in the world since it doesn't have any exposure in this world. Therefore, the course's view of AI and Yann's view of AI are completely opposite and different.