RACHAKONDA HRITHIK SAGAR	2023900021	PGSSP	ACTIVITY 5	
	Malicious use	**AI race**	**Organization risks**	**Rogue Ais**
Interpretability	Adversarial attacks, Hidden biases. Some examples are: financial fraud, social engineering	Exploiting Opacity, Hidden biases and for example: Disinformation campaigns	Decision-making bias, security vulnerabilities, develop and deploy interpretable AI models	User education and empowerment, addressing fairness issue and combating misinformation
Robustness	Data poisoning, and example is: spreading misinformation	exploiting weakness, data poisoning and examples i think can be: supply chain attacks	having sensor redundancy and security system, having realtime monitoring and threat detection	Data augmentation, Incorporating reasoning and commonsense
Deception detection	generating deepfakes and crafting synthetic identity	multi modal analysis, valid source verification and examples are: Adversarial Attack	utilizing inherently interpretable models, developing explainable frameworks and having a human in the loop	Identify manipulated content, trace the origin, build trust and user awareness, data filtering and validation, fact-checking
Monitoring	Face recognition, emotion detection and examples can be public awareness and education	Exploiting monitoring system, Data Harvesting and manipulation. examples are: Deep fakes	unfair and biases decisions, having a human in loop will be helpful	Focus on specific tasks and datatype, explainable methods, regular audits and evaluations, robust security measures
Unlearning	Remving biases that can lead to problems, examples can be: Removing data about how to suicide? kind of questions and adding mentalhelp details.	Data integrity and authentication, Storing data and examples are: unauthorised data removal	identify biases inputs, analyze fairness issues and have a score.	Biases traning data, limited diversity in training data, counterfactual training, fairness aware metrics and monitoring.