Course Type	Course Code	Name of Course	L	T	P	Credit
DE	OCSD606	Responsible & Safe AI Systems	3	0	0	3

Course Objective

There has been an exponential increase in the use of platforms / technologies like ChatGPT, Gemini, Llama, Sora, DALL-E, etc. in our day-to-day lives. These Language & Vision models have changed our way of living, and way we seek & create information. This course provides students with a comprehensive understanding of the ethical, social, and safety considerations essential for developing and deploying artificial intelligence (AI) systems.

Learning Outcomes

Uncover the intricacies of algorithmic transparency, fairness in machine (un)learning, interpretability, consistency and many more. The course encourages critical thinking and fosters a deep appreciation for the impact of AI on individuals and communities. Students who complete the course can: recognize possible harms that can be caused by modern AI capabilities; learn to reason about various perspectives on the trajectory of AI development and proliferation; learn about latest research agendas towards making AI systems safer.

Unit No.	Topics To be covered	No of Lecture Video	Video Time (Minutes)	Learning Outcome
1	Introduction, AI Capabilities Evolution Of AI, AI Risk Bias, Risks associated with AI getting harmful outputs from AI biases in AI models AI Risk Bias, AI Risk Bias. AI Capabilities Improvement in last 5-10 years. Imminent risks from AI Models: Toxicity, bias, goal misspecification, adversarial examples etc. Long-term risks from AI Models: Misuse, Misgeneralization, Rogue AGI. Principles of RAI - Transparency; Accountability; Safety, Robustness and Reliability; Privacy and Security; Fairness and non-discrimination; Human-Centred Values; Inclusive and Sustainable development, Interpretability	7	228	Evaluate the multi-year evolution of modern AI capabilities—including generative architectures and scalable foundation models—to understand both near-term breakthroughs and systemic safety implications. Analyze near-term and immediate AI model risks, including algorithmic bias, toxicity, adversarial vulnerabilities, and goal misspecification leading to harmful or skewed outputs. Assess long-term and existential AI risk profiles, spanning dual-use misuse scenarios, dangerous emergent capabilities, model misgeneralization, and rogue AGI alignment challenges. Implement Responsible AI (RAI) foundational principles—such as transparency, accountability, safety, reliability, fairness, and interpretability—to align model outputs with human-centered values. Formulate governance strategies that integrate privacy, security, non-discrimination, and sustainable development controls into the end-to-end AI lifecycle.
2	Robustness, RLHF, AI Alignment, Transformers, Introduction to Transformers architecture Encoder-Decoder BERT GPT Transformers. Recap of Deep Learning Techniques, Language/Vision Models. AI Risks for Gen models. Adversarial Attacks – Vision, NLP, Superhuman Go agents	8	268	Master the core mechanics of Transformer architectures, contrasting autoencoding (BERT) and autoregressive (GPT) models while surveying foundational deep learning techniques across language and vision domains. Evaluate AI alignment methodologies—including Reinforcement Learning from Human Feedback (RLHF)—to optimize model robustness, steerability, and safety against goal misspecification. Analyze security risks inherent to generative models, spanning hallucination vectors, data extraction vulnerabilities, and prompt injection mechanisms. Investigate cross-modal adversarial attacks—such as pixel perturbations in computer vision and typographic/token manipulations in NLP—to assess system fragile points. Examine adversarial policies engineered against superhuman Go AIs (e.g., KataGo), exposing blind spots in deep reinforcement learning and tree-search evaluation mechanisms.
3	Exact unlearning, approximate unlearning privacy unlearning concept, unlearning ask to forget, evaluation of unlearning WMDP TOFU Benchmark datasets Graph Unlearning, Representation engineering, Representation engineering Representation Engineering, Over fitting Under fitting, Feed-Forward Network Gradient Descent Loss Functions.	11	322	Master fundamental machine learning principles—including feed-forward networks, loss functions, gradient descent, non-linear classification, and data representation—while managing overfitting, underfitting, and model evaluation metrics. Execute exact, approximate, and graph-based machine unlearning protocols ("ask to forget") to uphold privacy rights, benchmarking performance against industry datasets like WMDP and TOFU.

	ML Poisoning Attacks like Trojans. Implications for current and future AI safety. Explainability Imminent and Long-term potential for transparency techniques Hands on, Evaluation metrics Data Preparation Data Representation. Building Classification Models Classification Non-linearity.			Analyze adversarial machine learning vulnerabilities, specifically data poisoning attacks, embedded model Trojans, and their broader implications for current and future AI safety. Apply representation engineering and explainability techniques to inspect internal activations, advance system transparency, and evaluate the imminent and long-term potential of AI interpretability tools.
4	Introduction to bias, Bias gender bias region bias religion bias, Source of Bias Source of bias, bias from data bias through guardrails, Mechanistic Interpretability Representation Engineering, model editing and probing Critiques of Transparency for AI Safety Hands on bias.	4	96	Categorize diverse sources and types of algorithmic bias—including gender, regional, and religious biases—tracing their roots from data collection through model guardrails and fine-tuning. Apply Mechanistic Interpretability techniques—such as internal state probing and representation engineering—to locate, analyze, and edit specific concepts within neural networks. Critically evaluate model editing methods and the limitations of transparency techniques in guaranteeing long-term AI safety and alignment. Conduct hands-on bias evaluations and interventions on machine learning models to detect, quantify, and mitigate unfair outputs across protected demographic attributes.
5	Bias stereo set crowS pairs Auto Debias, Coreference resolution Reddit Bias Context Gaurdrails Cobias, Detecting bias mitigating bias evaluation of bias, Deepen the understanding of Privacy & Fairness in AI	5	128	Benchmark and evaluate dataset-level stereotypes using tools like StereoSet and CrowS-Pairs, while executing automated mitigation algorithms such as AutoDebias to minimize social disparities. Identify contextual and relational biases in complex text using coreference resolution analysis across uncurated platforms like Reddit, while bypassing safety guardrails using constructed conversational probes like CoBia. Apply rigorous bias detection, context-oriented evaluation metrics, and debiasing techniques to advance fairness and privacy standards in artificial intelligence systems.
6	Data Privacy Privacy Utility Statistical privacy, Privacy Differential Privacy. Deepen the understanding of Privacy & Fairness in AI	2	120	Analyze the fundamental tradeoff between data privacy and statistical utility, optimizing data usability while maintaining robust statistical privacy guarantees. Implement formal Differential Privacy mechanisms—including privacy budget management (ϵ, δ) and noise addition—to prevent re-identification and inference attacks. Synthesize privacy-preserving techniques with algorithmic fairness frameworks to build secure, equitable, and privacy-aware machine learning models.
7	Approximate Differential Privacy, Exponential Mechanism, Fairness in Machine Learning Metrics and Tools for RAI - measuring bias/fairness, adversarial testing, explanations (Lime/SHAP/GradCam), audit mechanisms Regulation landscape - DPDP act (India), GDPR (EU), EU AI act, US presidential declaration, Ethical approvals, informed consent, participatory design, future of work, Indian context	3	123,	Evaluate advanced differential privacy mechanisms—including Approximate Differential Privacy (ϵ, δ) and the Exponential Mechanism—alongside fairness metrics to balance statistical utility with algorithmic non-discrimination. Execute Responsible AI (RAI) audits using bias/fairness measurement tools, adversarial testing, and post-hoc explainability techniques (LIME, SHAP, Grad-CAM). Analyze global and Indian regulatory frameworks—such as the DPDP Act, GDPR, EU AI Act, and US Executive Orders—while integrating ethical approvals, informed consent, and participatory design into the future of work.
8	Interpretability Transparency Probing, Codebook steering, What is AGI? When could it be achieved? Instrumental Convergence: Power Seeking, Deception etc.	4	129	Apply advanced interpretability and transparency techniques—such as probing and codebook steering—to inspect, control, and steer internal neural network representations. Evaluate technical definitions, architectural pathways, and timeline projections for achieving Artificial General Intelligence (AGI). Analyze emergent alignment risks driven by instrumental convergence—including power-seeking behavior, situational awareness, and strategic deception in advanced AI systems.
9	AI Policies Regulations AGI, AI Policies Regulations AGI Views on AGI Panel	3	143	Analyze evolving global policies, safety regulations, and expert panel perspectives surrounding Artificial General

	Discussion RAI in Legal domain RAI in Health care domain RAI in Education domain			Intelligence (AGI) timelines and existential risk governance. Apply Responsible AI (RAI) guardrails within legal systems to address accountability, algorithmic bias, IP governance, and liability in automated legal reasoning. Deploy domain-specific RAI frameworks across healthcare and education to ensure patient safety, diagnostic privacy, bias-free adaptive learning, and regulatory compliance.
10	AI Policies Regulations AGI A few other domains Policy issues in RAI	2	86	Analyze global AI policy frameworks, legislative trends, and regulatory governance models designed to mitigate systemic risks associated with Artificial General Intelligence (AGI). Examine cross-sector Responsible AI (RAI) deployments across diverse domains to ensure sector-specific ethical compliance, privacy, and operational safety. Address critical policy challenges in Responsible AI, including algorithmic accountability, intellectual property rights, liability allocation, and global compliance alignment.
11	Couple of panel discussion with industry practitioners, academic, government (possibly), and others. Finetuning Instruction-finetuning jailbreaking, AI governance, LLM Consistency.	3	100	Synthesize multi-stakeholder perspectives from industry, academia, and government to navigate real-world AI deployment, regulatory policy, and operational challenges. Execute full parameter fine-tuning and instruction-tuning methodologies to optimize Large Language Models for task-specific downstream applications. Analyze security exploits and jailbreaking vectors in LLMs to formulate robust AI governance structures, red-teaming protocols, and safety guardrails.
12	Fireside chat with eminent personalities Recorded Paper reading discussion Graph explanations GNN; s Higher order structures, Debiasing, Responsible AI, AI Capabilities Bias Interpretability Explainability RLHF Alignment AI risks	3	87	Engage with industry leaders through fireside chats and paper reading discussions to analyze cutting-edge research across Graph Neural Networks (GNNs), higher-order graph structures, and AI capabilities. Evaluate technical mechanisms in AI safety and alignment, spanning Reinforcement Learning from Human Feedback (RLHF), threat risk models, and advanced interpretability and explainability techniques. Formulate Responsible AI frameworks to detect and mitigate algorithmic bias, ensure graph-level fairness, and enforce ethical governance across complex machine learning architectures.
	Total No of video lectures & Video time	**55**	**1830**	

Text books

No Textbook available on the course topic as yet. All material will be made available via the NPTEL course website

Reference books:

No Reference book available on the course topic as yet. All material will be made available via the NPTEL course website

Link of NPTEL Course:

https://nptel.ac.in/courses/106106472