Indian Institute of Technology Kanpur

Proposal For a New Course

PHI44[illegible] (441)

1. Course Number: PHI 4XX 441
2. Title: Ethics of Artificial Intelligence
3. Per Week Lectures: 3(L), Tutorial: 0 (T), Laboratory: 0 (P), Additional Hours[0-2]: 0 (A),

   Credits (3*L+2*T+P+A): 9 Duration of Course: Full Semester
4. Proposing Department: Humanities and Social Sciences

   Other Departments which may be interested in the proposed course: N/A

   Other faculty members interested in teaching the proposed course: N/A
5. Proposing Instructor: Sushruth Ravish
6. Course Description: Artificial Intelligence is shaping society in profound ways, raising complex ethical dilemmas at both societal and individual levels. This course will philosophically explore the issues in the ethics of AI, focusing on issues such as fairness, transparency, trust, privacy, accountability, and societal impact. Students will learn to critically evaluate ethical frameworks and dilemmas in AI across diverse domains, including healthcare, autonomous systems, and social media.

   a. Objectives:

      i. Demonstrate an understanding of ethics and its application to AI.

      ii. Analyze ethical dilemmas in AI using tools from normative ethics, metaethics, and moral epistemology.

      iii. Develop a capacity to critically evaluate fairness, transparency, and accountability in AI systems.

      iv. Explore the societal, cultural, and policy implications of AI across domains.

      v. Learn how ethical principles can be integrated into the design of AI systems through practical frameworks and methodologies.

   b. Contents:

S.No	Broad Title	Topics	No. of Lectures
1	**Introduction to AI and Ethical Implications:**	Overview of AI, generative AI vs predictive AI, ethical concerns in AI development, societal impacts of AI, and ethical decision-making frameworks in AI systems.	3
2	**Foundations of Generative AI**	Understanding generative models like GANs, VAEs, and Diffusion models; Key concepts in unsupervised learning, ethical issues in synthetic data generation (e.g., deepfakes, authenticity).	3
3	**Predictive AI Models and Their Implications**	Introduction to supervised learning, predictive analytics, algorithmic bias, fairness in predictive modelling, and consequences of biased predictions in decision-making.	3
4	**Ethics of Data in AI**	Privacy issues, data ownership, consent in data usage, differential privacy, transparency in data collection, ethical data handling.	4
5	**AI and Accountability**	Black-box problem, Algorithmic accountability, responsibility in decision-making by AI, case studies of AI failures (e.g., bias in predictive policing, hiring algorithms).	4
6	**AI Alignment Problem**	Challenges in Embedding values in AI design, Possibility of moral machines	4
7	**Ethics of Algorithmic Decision-Making**	Ethics of predictive AI in areas like finance, criminal justice, and employment, analysing fairness, transparency, and the accountability of algorithmic decisions.	4
8	**AI in Autonomous Systems**	Ethical challenges in self-driving cars, drones, and autonomous robots, the trolley problem, moral decision-making in AI systems, responsibility and accountability in	4

		autonomous systems.	
9	**Studying Specific AI Issues and Controversies**	In-depth study of specific ethical issues (e.g., AI in military applications, surveillance technologies, AI in creative industries), regulatory challenges, and developing ethical guidelines.	5
10	**Future Directions and Governance of AI**	Ethical implications of emerging AI technologies, the role of global governance in AI regulation, transparency, and the need for ethical AI standards, interdisciplinary collaboration in AI ethics.	5

c. Pre-requisites, if any:Basic familiarity with Machine Learning is desirable

d. Short summary for including in the Courses of Study Booklet

This course will philosophically explore the issues in the ethics of AI, focusing on issues such as fairness, transparency, trust, privacy, accountability, and societal impact. Students will learn to critically evaluate ethical frameworks and dilemmas in AI across diverse domains, including healthcare, autonomous systems, and social media. Case studies and real-world applications will be used to connect theoretical concepts to practical challenges in AI ethics.

7. **Recommended Books:**

Textbooks:

i. *AI Ethics* (2023) by Paula Boddington. Springer.

ii. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities* (2023) by Luciano Floridi. OUP

References:

i. Negnevitsky, M. (2005). *Artificial Intelligence: A Guide to Intelligent Systems*. Addison Wesley.

ii. Kearns, M., & Roth, A. (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press.

iii. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

iv. O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing.

v. Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine Ethics*. Cambridge University Press.

vi. Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.

vii. Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

viii. Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. W.W. Norton & Company.

ix. Müller, V. C. (Ed.). (2020). *Ethics of Artificial Intelligence and Robotics*. Springer.

x. Van Wynsberghe, A. (2021). *Ethics in AI: A Practical Guide to Responsible AI*. Springer.

8. Any other remarks:

Sushruth R

Dated: 12/01/2025 Proposer (Sushruth Ravish)

Suchitra Mathur

Dated: 17/01/2025 DUGC Convener

The course is approved / not approved

Chairman, SUGC/SPGC

Dated: 6/3/2025