Training Course on Ethical AI in Generative Models

Professional Development Training Course Outline

Category	Data Science	**Duration**	5 Days
Base Fee	$1,500 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

The proliferation of **Generative AI** and **Large Language Models (LLMs)** has revolutionized industries, offering unprecedented capabilities in content creation, automation, and innovation. However, this transformative power comes with inherent **ethical challenges** related to **bias, fairness, accountability, transparency, and safety**. Organizations deploying these advanced AI systems face increasing scrutiny to ensure responsible development and deployment, mitigating risks such as discriminatory outcomes, misinformation, and intellectual property violations. This comprehensive training course provides essential knowledge and practical strategies to navigate the complex ethical landscape of **AI ethics** in **generative models**.

Training Course on Ethical AI in Generative Models: Addressing Bias, Fairness, and Safety in LLMs and Generative AI is designed to equip participants with a deep understanding of **responsible AI** principles, practical techniques for **bias detection and mitigation**, and robust frameworks for ensuring **AI safety** in real-world applications. Through a blend of theoretical insights, hands-on exercises, and pertinent **case studies**, attendees will learn to identify, assess, and address ethical considerations throughout the AI lifecycle, from data collection and model training to deployment and monitoring. By fostering a culture of **ethical AI development**, this course empowers professionals to build trustworthy, equitable, and impactful generative AI solutions that benefit society while upholding organizational values and regulatory compliance.

Programme Curriculum

Training Course on Ethical AI in Generative Models: Addressing Bias, Fairness, and Safety in LLMs and Generative AI

Introduction

The proliferation of **Generative AI** and **Large Language Models (LLMs)** has revolutionized industries, offering unprecedented capabilities in content creation, automation, and innovation. However, this transformative power comes with inherent **ethical challenges** related to **bias,**

fairness, accountability, transparency, and safety. Organizations deploying these advanced AI systems face increasing scrutiny to ensure responsible development and deployment, mitigating risks such as discriminatory outcomes, misinformation, and intellectual property violations. This comprehensive training course provides essential knowledge and practical strategies to navigate the complex ethical landscape of **AI ethics** in **generative models**.

Training Course on Ethical AI in Generative Models: Addressing Bias, Fairness, and Safety in LLMs and Generative AI is designed to equip participants with a deep understanding of **responsible AI** principles, practical techniques for **bias detection and mitigation**, and robust frameworks for ensuring **AI safety** in real-world applications. Through a blend of theoretical insights, hands-on exercises, and pertinent **case studies**, attendees will learn to identify, assess, and address ethical considerations throughout the AI lifecycle, from data collection and model training to deployment and monitoring. By fostering a culture of **ethical AI development**, this course empowers professionals to build trustworthy, equitable, and impactful generative AI solutions that benefit society while upholding organizational values and regulatory compliance.

Course Duration

5 days

Course Objectives

1. Understand the core principles of **AI ethics**, including fairness, transparency, accountability, and privacy in the context of generative models.
2. Recognize various forms of **algorithmic bias** in **LLMs** and other generative AI systems, stemming from training data, model architecture, and deployment contexts.
3. Learn and apply advanced techniques for **bias detection** and **de-biasing strategies** in generative AI pipelines.
4. Develop and evaluate generative models for **fairness metrics**, understanding different fairness definitions and their implications.
5. Grasp the critical aspects of **AI safety**, including guardrails, robust alignment, and preventing unintended harmful outputs from generative AI.
6. Understand best practices for **responsible data collection**, data privacy, and secure data handling for training **generative AI** models.
7. Explore techniques for **model interpretability** and explainability in LLMs, enhancing transparency and trust.
8. Analyze the risks of **AI-generated misinformation** and deepfakes, and learn strategies for detection and prevention.
9. Familiarize with emerging global **AI regulations** and ethical guidelines (e.g., GDPR, EU AI Act) relevant to generative AI.
10. Construct and apply practical **ethical AI frameworks** for responsible innovation within organizations.
11. Learn to perform **AI ethics impact assessments** to proactively identify and mitigate potential societal and ethical risks.
12. Understand the considerations for **responsible AI deployment** and continuous monitoring of generative models in production.
13. Cultivate a strong understanding of the societal implications and **ethical considerations** of generative AI among stakeholders.

Organizational Benefits

- Reduce legal, reputational, and financial risks associated with biased or unsafe AI systems, ensuring adherence to evolving **AI regulations** and **data privacy** laws.
- Build public and stakeholder trust by demonstrating a commitment to **responsible AI development** and ethical innovation, strengthening brand image.
- Develop more robust, fair, and reliable generative AI applications, leading to better business outcomes and enhanced operational efficiency.
- Foster a culture of **ethical innovation**, attracting top talent and enabling the responsible adoption of cutting-edge **generative AI technologies**.
- Contribute positively to society by proactively addressing issues of **bias, fairness, and accountability**, aligning with corporate social responsibility goals.
- Establish internal mechanisms for identifying, assessing, and mitigating **AI ethics** challenges throughout the AI lifecycle.
- Minimize costly post-deployment fixes by integrating ethical considerations early in the **AI development lifecycle**.

Target Audience

1. AI/ML Engineers & Data Scientists
2. Product Managers & AI Project Leads.
3. AI Ethicists & Researchers
4. Legal & Compliance Professionals
5. Business Leaders & Executives
6. Policy Makers & Regulators
7. UX/UI Designers
8. Auditors & Governance Professionals.

Course Outline

Module 1: Foundations of Ethical Generative AI

- Defining **Ethical AI** in the era of generative models: principles and values.
- Introduction to **Generative Models** and **LLMs**: architecture and capabilities.
- Understanding the societal impact and dual-use potential of advanced AI.
- Key ethical considerations: **fairness, accountability, transparency, privacy, safety**.
- Historical context and evolution of **AI ethics** discussions.
- **Case Study:** The early days of facial recognition bias in consumer products, highlighting the need for proactive ethical considerations.

Module 2: Unpacking AI Bias in Generative Models

- Types of **algorithmic bias**: dataset bias, interaction bias, aggregation bias, and systemic bias.
- Sources of bias in **LLMs** and generative AI: training data, pre-training, fine-tuning.
- Quantifying and identifying bias: statistical methods and qualitative analysis.
- Bias in text generation: gender stereotypes, racial bias, and offensive content.
- Bias in image generation: representation, stereotypes, and cultural insensitivity.
- **Case Study:** Generative AI models perpetuating gender stereotypes in job descriptions or image generation (e.g., "doctor" often generates male images).

Module 3: Strategies for Bias Detection and Mitigation

- Data-centric approaches: **data augmentation**, balanced datasets, and **de-biasing** techniques for training data.
- Model-centric approaches: architectural adjustments, regularization, and adversarial training.

- Post-processing techniques: re-ranking, calibration, and output filtering.
- Human-in-the-Loop (HITL) strategies for continuous bias monitoring.
- Tools and frameworks for **bias detection** and mitigation (e.g., Fairlearn, AI Fairness 360).
- **Case Study:** A company successfully de-biasing a hiring LLM by using diverse data augmentation and re-weighting techniques, leading to more equitable candidate selection.

Module 4: Ensuring Algorithmic Fairness in Practice

- Different definitions of **fairness**: demographic parity, equal opportunity, individual fairness.
- Evaluating fairness metrics for generative model outputs.
- Trade-offs between **fairness**, accuracy, and utility in AI systems.
- Fairness-aware model design and optimization.
- Implementing **fairness guardrails** in generative AI deployment.
- **Case Study:** Analyzing a loan approval generative AI system that inadvertently showed disparate impact on minority groups, and how a fairness-aware approach corrected the issue.

Module 5: AI Safety and Robustness in LLMs

- Defining **AI safety**: alignment, harmful outputs, and unintended consequences.
- Techniques for **robustness**: adversarial attacks and defenses for generative models.
- Preventing **hallucinations** and factual inaccuracies in **LLM** outputs.
- Developing **safety mechanisms** and content filters for generative AI.
- Red teaming and stress testing generative models for safety vulnerabilities.
- **Case Study:** An LLM generating harmful or biased recommendations, despite being trained on seemingly benign data, highlighting the need for robust safety testing.

Module 6: Data Governance and Privacy in Generative AI

- **Data ethics**: consent, data provenance, and intellectual property considerations.
- GDPR and other **data privacy** regulations relevant to generative AI training data.
- Secure data handling, anonymization, and differential privacy techniques.
- Managing synthetic data and its implications for privacy.
- Establishing clear **data governance** policies for generative AI projects.
- **Case Study:** A generative model inadvertently reproducing sensitive personal data from its training set, emphasizing the importance of strong privacy safeguards.

Module 7: Transparency, Interpretability, and Accountability (XAI)

- The "black box" problem in deep learning and **LLMs**.
- **Explainable AI (XAI)** techniques: LIME, SHAP, attention mechanisms.
- Interpreting generative model outputs and decision-making processes.
- Establishing clear lines of **accountability** for AI system behavior.
- Documentation and reporting standards for **transparent AI systems**.
- **Case Study:** Using XAI tools to understand why a generative model produced a particular, unexpected output, leading to insights for model improvement and increased trust.

Module 8: Ethical Deployment and Continuous Monitoring

- Frameworks for **responsible AI deployment** and MLOps.
- Continuous monitoring of **generative AI** systems for emerging biases and safety issues.
- Developing feedback loops and user reporting mechanisms for ethical concerns.
- The role of human oversight in complex **AI systems**.
- Building an **ethical AI culture** within the organization.

- **Case Study:** A generative AI chatbot deployed for customer service starts generating problematic responses over time, leading to a system for continuous monitoring and rapid intervention.

Training Methodology

This course employs a **blended learning approach**, combining:

- **Interactive Lectures**: Engaging presentations with clear explanations of complex concepts.
- **Hands-on Workshops**: Practical exercises using industry-standard tools and frameworks for **bias detection, mitigation, and fairness evaluation**.
- **Real-world Case Studies**: In-depth analysis of actual ethical challenges and solutions in **generative AI** and **LLMs**.
- **Group Discussions & Debates**: Fostering critical thinking and diverse perspectives on **AI ethics dilemmas**.
- **Live Demos**: Showcasing practical applications of ethical AI principles and tools.
- **Guest Speakers**: Industry experts sharing their insights and experiences in **responsible AI**.
- **Q&A Sessions**: Dedicated time for participants to ask questions and clarify doubts.

Register as a group from 3 participants for a Discount

Send us an email: info@fineskilltrainingcenter.org or call *+254769199797*

Certification

Upon successful completion of this training, participants will be issued with a globally- recognized certificate.

Tailor-Made Course

We also offer tailor-made courses based on your needs.

Key Notes

a. The participant must be conversant with English.

b. Upon completion of training the participant will be issued with an Authorized Training Certificate

c. Course duration is flexible and the contents can be modified to fit any number of days.

d. The course fee includes facilitation training materials, 2 coffee breaks, buffet lunch and A Certificate upon successful completion of Training.

e. One-year post-training support Consultation and Coaching provided after the course.

f. Payment should be done at least a week before commence of the training, to Fineskill Training Center account, as indicated in the invoice so as to enable us prepare better for you.

Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
23 Nov 2026	27 Nov 2026	Online	$1,000
23 Nov 2026	27 Nov 2026	Physical	$1,500
23 Nov 2026	27 Nov 2026	Physical	$3,000
23 Nov 2026	27 Nov 2026	Physical	$3,000
23 Nov 2026	27 Nov 2026	Physical	$6,000
23 Nov 2026	27 Nov 2026	Physical	$4,500
23 Nov 2026	27 Nov 2026	Physical	$6,000
23 Nov 2026	27 Nov 2026	Physical	$6,000

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now

Email: info@fineskilltrainingcenter.com | Website: www.fineskilltrainingcenter.com