COURSE SPECIFICATION DOCUMENT

Academic School / Department: Science, Innovation & Technology

Programme: MSc Artificial Intelligence

***FHEQ* Level:** 7

Course Title: AI Security

Course Code: COMP 7105

Total Hours: 200 (Lev 7) (4 US Credits)
Timetabled Hours: 39
Guided Learning Hours 21
Independent Learning Hours: 140

Credit 20 UK CATS credits
10 ECTS credits
4 US credits

Course Description:
AI Security is focused on the protection, robustness, and trustworthy operation of artificial intelligence systems. The module examines the vulnerabilities inherent in modern machine learning architectures, the techniques used to attack them, and the defences required to secure AI in real-world applications. Students study adversarial machine learning, model extraction, data poisoning, backdoor attacks, prompt-based LLM exploitation, privacy attacks, to name a few. The module integrates both theoretical perspectives and practical methodologies enabling students to evaluate and implement security controls across the AI lifecycle.

Prerequisites:
None

Aims and Objectives:
The aims of this module are to develop advanced understanding and practical competence in securing AI systems.
To gain insights into the vulnerabilities present across the AI development and deployment pipeline, and to learn techniques for defending against adversarial and systemic threats.

Objectives:

- Analyse and evaluate adversarial attacks, data poisoning, backdoors, model extraction and other AI-targeted threats.
- Apply practical techniques for assessing robustness and conducting AI security testing.

- Implement defensive methods such as adversarial training, input sanitisation, monitoring and model hardening.
- Design secure AI workflows and incorporate security-by-design principles into model development and deployment.

Programme Outcomes:
A1, A3, B2, B3, C1, C3, D3.

A detailed list of the programme outcomes are found in the Programme Specification. This is located at the archive maintained by Registry and found at:
https://www.richmond.ac.uk/programme-and-course-specifications/

Learning Outcomes:
By the end of this course, students will be able to:

1. Demonstrate advanced understanding of AI security concepts, threat models, and the vulnerabilities affecting machine learning and LLM-based systems.
2. Critically analyse and evaluate adversarial attacks, data poisoning, backdoors, model extraction techniques, and other model-targeted threats.
3. Design and conduct AI security testing, including robustness evaluation, adversarial experimentation and diagnostic analysis of model behaviour.
4. Implement defence strategies such as adversarial training, input filtering, detection methods, monitoring systems, and secure deployment practices.
5. Communicate technical findings and security evaluations clearly, professionally and in accordance with postgraduate academic standards.

Indicative Content:

- Model vulnerabilities
- Adversarial attacks
- Data poisoning
- Backdoor attacks
- Model extraction
- Robustness evaluation
- Adversarial training
- Input sanitisation
- Detection methods
- Monitoring systems
- Secure deployment
- LLM security issues

Assessment:
This course conforms to the University Assessment Norms approved at Academic Board and located at: https://www.richmond.ac.uk/university-policies/

Teaching Methodology:

Teaching includes lectures, practical laboratory sessions, guest lectures, and guided learning activities.

Indicative Text(s):

- Huang, K., Wang, Y., Goertzel, B., Li, Y., Wright, S. and Ponnapalli, J. (eds.) (2024). *Generative AI security: theories and practices*. Cham: Springer Nature Switzerland.
- Sotiropoulos, J. (2024). *Adversarial AI Attacks, Mitigations, and Defense Strategies: A cybersecurity professional's guide to AI attacks, threat modeling, and securing AI with MLSecOps*. Birmingham: Packt Publishing Ltd.
- Wilson, S. (2024). *The Developer's Playbook for Large Language Model Security*. Farnham: O'Reilly.

Journals

- *Applied Artificial Intelligence.*
- *Journal of experimental and theoretical artificial intelligence.*

Websites

- AISafety.info *AI Safety Information and Resources*. Available at: https://aisafety.info (Accessed: December 2025). Focuses on AI safety, risk, alignment, and governance.
- RobustBench (2025) *RobustBench: Adversarial Robustness Benchmark*. Available at: https://robustbench.github.io (Accessed: December 2025). A standardised benchmark for evaluating adversarial robustness of models.
- Hugging Face. Available at: https://huggingface.co/docs (Accessed: December 2025). Official documentation hub for Hugging Face's machine learning tools and libraries.
- DeepMind (2025) *Responsibility & Safety*. Available at: https://deepmind.google/responsibility-and-safety (Accessed: December 2025). DeepMind's overview of its AI safety and responsibility work.

See syllabus for complete reading list.

Change Log for this CSD:

Nature of Change	Date Approved & Approval Body (School or AB)	Change Actioned by Registry Services
Guided Learning Hours menu updated	October 2025	
Total Hours Updated	October 2025	
