COURSE UNIT DESCRIPTION

Course unit title	Course unit code
Artificial Intelligence Security	**ITAIS**

Lecturer	Department where the course unit is delivered
Coordinator: Teach. Asst. Virgilijus Krinickij	Department of Computational and Data Modeling Faculty of Mathematics and Informatics Vilnius University

Cycle	Type of the course unit
First	Individual

Mode of delivery	Semester or period when the course unit is delivered	Language of instruction
Face-to-face	Autumn	Lithuanian, English

Prerequisites
General understanding of Machine Learning. Proficiency in Python programming. Cybersecurity fundamentals.

Number of ECTS credits allocated	Student's workload	Contact hours	Individual work
5	142	66	76

Purpose of the course unit: programme competences to be developed

Generic competences to be developed

- Ability to apply knowledge in practical situations *(BK1)*
- Knowledge and understanding of the subject area and understanding of the profession *(BK2)*
- Ability for abstract thinking, processing and analysing information *(BK3)*
- Ability to use information and communication technologies *(BK5)*

Subject-specific competences to be developed

- Ability to apply general methods of the program design, make and analyse software requirements *(DK1)*
- Ability to analyse the algorithmic process of the task based on the general properties of the algorithm *(DK2)*
- Ability to do program and IT service testing and debugging *(DK4)*
- Ability to ensure information security using management and security mechanisms of operating systems and software *(DK8)*

Learning outcomes of the course unit	Teaching and learning methods	Assessment methods
Explain the architecture and operational principles of machine learning models, LLMs, and agentic AI systems.	Lectures with examples on Conceptual foundations (AI, ML, LLMs, adversarial ML). Security frameworks and threat models.	Assessment is designed to evaluate both theoretical understanding and practical capability during the project presentation**.**
Identify and categorize security threats across the AI lifecycle **(**data, model, inference, deployment**).**	Lectures and labs on securing LLM pipelines (RAG, APIs, tools). Reading technical material.	Practical tasks.
Describe key adversarial attack techniques, including poisoning, evasion, and model extraction.	Labs on adversarial ML (poisoning, evasion).	Practical tasks.
Understand LLM-specific vulnerabilities such as prompt injection, jailbreaks, and data leakage.	Labs on prompt injection and jailbreak. Experimentation with tools (e.g., open-source models, frameworks).	Practical tasks.

Course content: breakdown of the topics	Individual work: time and assignments							
	Lectures	Tutorials	Seminars	Laboratory work	*Consultations during LW*	**Contact hours**	**Individual work**	**Assignments**
1. Introduction to Adversarial AI	2			2			**2**	Laboratory work
2. Machine Learning Attack Surface	2			2			**2**	Laboratory work
3. Data Poisoning Attacks	2			2			**2**	Laboratory work
4. Backdoor & Trojan Attacks	2			2			**5**	Laboratory work
5. Evasion Attacks	2			2			**5**	Laboratory work
6. Model Extraction & Inference Attacks	2			2			**4**	Laboratory work
7. Defenses Against Adversarial Attacks	2			2			**5**	Laboratory work
8. LLM Fundamentals for Security	2			2			**4**	Laboratory work
9. LLM Threat Model	2			2			**5**	Laboratory work
10. Prompt Injection Attacks	2			2			**5**	Laboratory work
11. Jailbreaking & Safety Bypass	2			2			**4**	Laboratory work
12. Securing LLM Applications	2			2			**4**	Laboratory work
13. RAG & Tooling Security	2			2			**5**	Laboratory work
14. Agentic AI Fundamentals	2			2			**5**	Laboratory work
15. Offensive Agent Design	2			2			**5**	Laboratory work
16. Agent vs Agent Security & Automation	2			2			**4**	Laboratory work
Self-preparation and exam.							**10**	Individual study of literature
Total	**32**			**32**		**66**	**76**	

Assessment strategy	Weight %	Deadline	Assessment criteria
Exam	50%	At the end of the semester.	Questions that assess the ability to apply the gained theoretical and practical knowledge.
Project	50%	At the end of the semester.	Project defense. Assessment criteria: ability to discuss the topic of the task, answer questions, and purposeful completion of the task.
Course retake externally			The course cannot be retaken externally.

Author	Publishing year	Title	Issue No or volume	Publishing house or Internet site
Required reading				
Steve Wilson	2024	The Developer's Playbook for Large Language Model Security: Building Secure AI Applications	1st Edition	O`Reilly
John Sotiropoulos	2026	Adversarial AI Attacks, Mitigations, and Defense Strategies: A cybersecurity professional's guide to AI attacks, threat modeling, and securing AI with MLSecOps		Packt
Orhan Yildirim	2026	Agentic AI for Offensive Cybersecurity		Packt

DALYKO APRAŠAS

Dalyko pavadinimas	Kodas
Dirbtinio intelekto saugumas	**ITDIS**

Dėstytojas	Padalinys
Koordinuojantis: Jaunesnysis asistentas Virgilijus Krinickij	Kompiuterinio ir duomenų modeliavimo katedra Matematikos ir informatikos fakultetas Vilniaus universitetas

Studijų pakopa	Dalyko tipas
Pirmoji	Pasirenkamasis

Įgyvendinimo forma	Vykdymo laikotarpis	Vykdymo kalba
Auditorinis	Rudens	Lietuvių, anglų

Reikalavimai studijuojančiajam
Bendras mašininio mokymosi supratimas. Python programavimo įgūdžiai. Kibernetinio saugumo pagrindai.

Dalyko apimtis kreditais	Visas studento darbo krūvis	Kontaktinio darbo valandos	Savarankiško darbo valandos
5	142	66	76

Dalyko tikslas: studijų programos ugdomos kompetencijos

Bendrosios kompetencijos:

- Gebėjimas taikyti žinias praktinėse situacijose *(BK1)*
- įgyti dalykinės srities žinių ir suprasti savo profesiją *(BK2)*,
- Gebėjimas abstrakčiai mąstyti, apdoroti ir analizuoti informaciją (*BK3*),
- įgyti informacinių bei komunikacijos technologijų naudojimo patirties *(BK5)*,

Dalykinės kompetencijos

- taikyti programų projektavimo bendruosius metodus, formuluoti ir analizuoti programinės įrangos reikalavimus *(DK1)*,

- Gebėjimas analizuoti užduoties algoritminį procesą remiantis bendromis algoritmo savybėmis *(DK2)*
- testuoti ir derinti programas ir IT paslaugas *(DK4)*,
- Gebėjimas užtikrinti informacijos saugumą naudojant operacinių sistemų ir programinės įrangos valdymo ir saugumo mechanizmus *(DK8)*.

Dalyko studijų siekiniai	Studijų metodai	Vertinimo metodai
Paaiškinti mašininio mokymosi modelių, LLM ir agentinių dirbtinio intelekto sistemų architektūrą ir veikimo principus.	Paskaitos su pavyzdžiais apie konceptualius pagrindus (DI, ML, LLM, priešiškas ML). Saugumo sistemos ir grėsmių modeliai.	Vertinimas tiek teorinis supratimas, tiek praktinius gebėjimus projekto pristatymo metu.
Nustatyti ir suskirstykite saugumo grėsmes į kategorijas visame dirbtinio intelekto gyvavimo cikle (duomenys, modelis, išvados, diegimas).	Paskaitos ir laboratoriniai darbai apie LLM kanalų apsaugą (RAG, API, įrankiai). Techninės medžiagos skaitymas.	Praktinės užduotys.
Apibūdinti pagrindinius priešiškųjų atakų metodus, įskaitant užklausų injekcijas, apėjimas ir modelio išgavimą.	Laboratoriniai darbai apie priešiškus ML (apnuodijimas, apėjimas).	Praktinės užduotys.
Suprasti LLM būdingus pažeidžiamumus, tokius kaip užklausų injekcijos, nulaužimas ir duomenų nutekėjimas.	Laboratoriniai darbai apie užklausų injekcijas ir modelio vogimus. Eksperimentavimas su įrankiais (pvz., atvirojo kodo modeliais, karkasais).	Praktinės užduotys.

Temos	**Kontaktinio darbo valandos**							**Savarankiškų studijų laikas ir užduotys**	
	Paskaitos	Konsultacijos	Seminarai	Pratybos	Laboratoriniai darbai (LD)	*Konsultavimas LD metu*	**Visas kontaktinis darbas**	**Savarankiškas darbas**	**Užduotys**
1. Įvadas į priešišką dirbtinį intelektą (DI)	2			2				**2**	Virtualizacijos sprendimų pavyzdžiai
2. Mašininio mokymosi atakų paviršius	2			2				**2**	Laboratorinis darbas
3. Užklausų užkrėtimo atakos	2			2				**2**	Laboratorinis darbas
4. Galinių durų ir Trojos arklių atakos	2			2				**5**	Laboratorinis darbas
5. Apėjimo atakos	2			2				**5**	Laboratorinis darbas
6. Modelio išgavimo ir samprotavimo atakos	2			2				**4**	Laboratorinis darbas
7. Apsauga nuo nuotolinių atakų	2			2				**5**	Laboratorinis darbas
8. LLM saugumo pagrindai	2			2				**4**	Laboratorinis darbas
9. LLM grėsmių modelis	2			2				**5**	Laboratorinis darbas
10. Užklausų atakos	2			2				**5**	Laboratorinis darbas
11. Įsilaužimas ir saugumo apėjimas	2			2				**4**	Laboratorinis darbas

12. LLM programų apsauga	2			2				**4**	Laboratorinis darbas
13. RAG ir įrankių saugumas	2			2				**5**	Laboratorinis darbas
14. Agentinio DI pagrindai	2			2				**5**	Laboratorinis darbas
15.Atakos pasinaudojus agentais	2			2				**5**	Laboratorinis darbas
16. Agentų saugumas ir automatizavimas	2			2				**4**	Laboratorinis darbas
Pasiruošimas egzaminui ir jo laikymas.								**10**	Savarankiškas literatūros studijavimas
Iš viso	**32**			**32**			**66**	**76**	

Vertinimo strategija	**Svoris proc.**	**Atsiskaitymo laikas**	**Vertinimo kriterijai**
Egzaminas	50%	Semestro pabaigoje	Egzamino klausimai įvertinti teorinių ir praktinių žinių taikymą.
Projektas	50%	Atsiskaitymas semestro gale	Projekto gynimas. Vertinimo kriterijai: gebėjimas diskutuoti užduoties tema, atsakyti į klausimus, tikslingas užduoties atlikimimas.
Kurso perlaikymas eksternu			Kursas negali būti perlaikytas eksternu.

Autorius	Leidi-mo metai	Pavadinimas	Periodinio leidinio Nr. ar leidinio tomas	Leidimo vieta ir leidykla ar internetinė nuoroda
Required reading				
Steve Wilson	2024	The Developer's Playbook for Large Language Model Security: Building Secure AI Applications	1st Edition	O`Reilly
John Sotiropoulos	2026	Adversarial AI Attacks, Mitigations, and Defense Strategies: A cybersecurity professional's guide to AI attacks, threat modeling, and securing AI with MLSecOps		Packt
Orhan Yildirim	2026	Agentic AI for Offensive Cybersecurity		Packt