Title: Secure Agentic AI Systems
Long Title: Secure Agentic AI Systems
Credits: 5
NFQ Level: Expert

Module Description:

This module introduces learners to the principles, architectures, security considerations, and governance of agentic AI systems. Agentic AI refers to systems composed of autonomous or semi-autonomous agents capable of goal-directed behaviour, interaction, and adaptation within dynamic environments. While such systems offer significant potential, they also expand digital attack surfaces and raise challenges relating to control, exploitability, explainability, assurance, and accountability.
The module emphasises the design of bounded and governed agentic systems operating in adversarial contexts. Learners critically evaluate agent behaviour, autonomy boundaries, coordination mechanisms, and human oversight models, while applying structured threat modelling, risk assessment, and assurance strategies. Particular attention is given to secure architectural design, adversarial failure modes, and the proportional governance of autonomous systems consistent with responsible and secure AI practice.

Learning Outcomes

On successful completion of this module the learner will be able to:

- **LO1** Critically analyse agentic AI systems and distinguish between different forms of autonomy, coordination, and human oversight in adversarial digital environments.
- **LO2** Design and implement proof-of-concept bounded agentic AI systems that demonstrate goal-directed behaviour while maintaining transparency and secure controls.
- **LO3** Evaluate risks, threat models, assumptions, and failure modes inherent in agentic AI systems, including misuse of system capabilities.
- **LO4** Apply governance, cybersecurity assurance, and testing strategies to ensure agentic AI systems operate safely, resiliently, accountably, and explainable.
- **LO5** Justify design and governance decisions using structured reasoning, empirical evaluation, and critical reflection.

Indicative Content

Foundations of Agentic AI and Autonomy
This section introduces the core concepts underpinning agentic AI systems, including perception, decision-making, action, and feedback within an environment. Learners explore different levels of autonomy and examine how goal-directed behaviour differs from conventional automation. The delegation of decision-making to autonomous systems is considered alongside issues of control, predictability, responsibility, and the expansion of digital attack surfaces in adversarial environments.

Architectures for Single-Agent and Multi-Agent Systems
This section examines architectural patterns for designing agentic systems in a framework-agnostic manner. Learners explore observe–decide–act cycles, modular and layered design, and structured interaction with environments. Multi-agent configurations are introduced with emphasis on coordination, communication, trust boundaries, and clearly defined operational and security constraints, including least-privilege tool access and bounded capability design.

Role-Based and Structured Agent Design
This section considers role separation within agentic systems, such as planning, execution, validation, and review functions. Learners explore how internal feedback loops, policy enforcement layers, and oversight mechanisms can improve transparency, reliability, and resilience against misuse or unintended behaviour.

Emergent Behaviour and Risk
This section addresses the risks associated with autonomous systems, including unintended optimisation, misalignment, emergent behaviour, and adversarial manipulation. Learners examine how system-level outcomes can diverge from design intent and consider structured risk assessment approaches, including threat modelling, in dynamic and potentially hostile environments.

Governance, Human Oversight, and Assurance
This section integrates governance into technical and security design thinking. Learners explore models of human-in-the-loop and human-in-command oversight, accountability mapping, escalation mechanisms, and documented operational boundaries. Risk-based governance approaches aligned with European regulatory principles are examined at a conceptual level. The emphasis is on applying assurance strategies that make systems controllable, reviewable, secure, and accountable within organisational contexts.

Testing, Traceability, and Justifiable Operation
This section focuses on how agentic systems can be tested, monitored, and understood in practice. Learners explore approaches to validation, adversarial testing, scenario-based assessment, and structured logging to ensure that system behaviour can be reviewed and, where necessary, investigated. The importance of explainability, traceability, and operational monitoring is considered in supporting transparency, stakeholder confidence, and ongoing oversight in security-sensitive environments.

Ethical and Regulatory Contexts for Deployment
This section situates agentic AI within broader societal, regulatory, and cybersecurity environments. Learners critically examine principles such as fairness, transparency, proportionality, accountability, and risk-tiered regulation. Emphasis is placed on articulating and defending design, security, and governance decisions in light of ethical considerations and organisational responsibilities.

Course Work

Assessment Type	*Assessment Description*	*Outcome Addressed*	*% of Total*	*Assessment Date*
Project	Assessment Description Critical Analysis of an Agentic AI System. Each student independently selects and critically analyses one real or emerging agentic AI system (commercial, open-source, research-based, multiagent, or a clearly defined hypothetical system) and justifies why it qualifies as "agentic." Using module theory, students must define agentic AI and analyse the system's type of autonomy, coordination model (if	1,3,5	40	Week 6

	relevant), level of human oversight, and whether it operates within clear operational and security bounds. They must examine key assumptions, dependencies, trust boundaries, and attack surfaces, and conduct a structured risk analysis addressing automation bias, misalignment, emergent behaviour, and adversarial manipulation (e.g., prompt injection or tool misuse). The assignment should critically evaluate the balance between autonomy, security, and governance, and conclude by reflecting on limitations, uncertainties, and residual risk. Findings should be fully documented and/or presented.			
Project	Designing and Assuring a Bounded Agentic AI System. Students are required to design and justify a bounded agentic AI system in a realistic domain. The submission must articulate the system's purpose, architectural structure, autonomy boundaries, human oversight model, and security controls. Students must include a structured risk and threat analysis addressing failure modes, adversarial risks, misalignment, and emergent behaviour. They must propose proportionate governance and assurance mechanisms (e.g., logging, testing, monitoring, access controls, auditability, regulatory alignment). The project should demonstrate integrative reasoning, critical judgement, awareness of residual risk, and secure-by-design thinking. Findings should be fully documented and/or presented.	1,2,3,4,5	60	Sem End

No End of Module Formal Exam

Assessment Breakdown	%
Coursework	100

Re-Assessment Requirement

Coursework Only

This module is reassessed solely on the basis of re-submitted coursework. There is no repeat written examination.

Workload – Full Time

Workload Type	*Workload Description*	*Hours*	*Frequency*	*Average Weekly Leaner Workload*
Lecture	Lectures covering the concepts underpinning the learning outcomes.	2	Every Week	2.00
Lab	Lab to support learning outcomes.	2	Every Week	2.00
Independent & Directed Learning (Non-contact)	Independent learning.	3	Every Week	3.00
		Total Hours		7.00
		Total Weekly Learner Workload		7.00
		Total Weekly Contact Hours		4.00

Workload – Part Time

Workload Type	*Workload Description*	*Hours*	*Frequency*	*Average Weekly Leaner Workload*
Lecture	Lectures covering the concepts underpinning the learning outcomes.	2	Every Week	2.00
Lab	Lab to support learning outcomes.	2	Every Week	2.00
Independent & Directed Learning (Non-contact)	Independent Study.	3	Every Week	3.00
		Total Hours		7.00
		Total Weekly Learner Workload		7.00
		Total Weekly Contact Hours		4.00

Recommended Resources

Pascal Bornet,Jochen Wirtz,Thomas H. Davenport,David De Cremer,Brian Evergreen,Phil Fersht,Rakesh Gohel,Shail Khiyara,Pooja Sund,Nandan Mullakara. (2025), Agentic Artificial

Intelligence: Harnessing AI Agents to Reinvent Business, work and Life, Irreplaceable Publishing, p.546, [ISBN: 979-8992833614].

Anjanava Biswas,Wrick Talukdar. (2025), Building Agentic AI Systems: Create intelligent, autonomous AI agents that can reason, plan, and adapt, Packt Publishing Ltd, p. 292, [ISBN: 978-1801079273].

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. (2021), On the dangers of stochastic parrots: Can language models be too big?, n Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, https://dl.acm.org/doi/10.1145/3442188.3445922

Park, Joon Sung, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. (2023), Generative agents: Interactive simulacra of human behavior., In Proceedings of the 36th annual acm symposium on user interface software and technology, https://dl.acm.org/doi/10.1145/3586183.3 606763

Murugesan, San. (2025), The rise of agentic AI: implications, concerns, and the path forward, IEEE Intelligent Systems, Volume 40, Issue 2, https://ieeexplore.ieee.org/abstract/doc ument/10962241

online, European Union. Searchable EU Publications database, EU, https://op.europa.eu/en/web/general-publ ications/publications

online, EU. The EU Artificial Intelligence Act Up-to-date developments and analyses of the EU AI Act, EU,

https://artificialintelligenceact.eu/

online, NIST. AI Risk Management Framework, NIST,

https://www.nist.gov/itl/ai-risk-managem ent-framework

online, MITRE Corporation. MITRE ATLAS, MITRE,

https://atlas.mitre.org/

online, OWASP. OWASP Top 10 for Large Language Model Applications, OWASP,

https://owasp.org/www-project-top-10-for -large-language-model-applications/

online, ENISA. (2023), Artificial Intelligence and Cybersecurity Research, ENISA,

https://www.enisa.europa.eu/publications /artificial-intelligence-and-cybersecuri ty-research