Cyber Security Project Proposals: Semester 2, 2025

P01_PROJECT DETAILS	
Project Title	Evaluating the effectiveness of DP in protecting GNNs
Project Description	With the increasing deployment of Graph Neural Networks (GNNs) in privacy-sensitive applications such as social networks, healthcare, and financial systems, ensuring the privacy of individuals represented in graph data has become critical. Differential Privacy offers formal privacy guarantees and has been applied to GNN training as a defence mechanism. However, the actual strength of DP protection in GNNs against modern attack techniques remains insufficiently understood. This project aims to systematically evaluate the effectiveness of DP in defending GNNs against privacy attacks that exploit the graph structure. The focus is to understand how much protection is truly provided, and under what circumstances sensitive information may still be leaked.
Project Skills	☒ AI/Machine Learning ☒ Cyber Security ☒ Programming (Software, Mobile and Web Development)
Environment	Kotlin, Python
Research Component	This project investigates the effectiveness of DP in GNNs against various privacy threats specific to graph-structured data. While DP has been widely applied in traditional machine learning to protect individual data points, its protective power in the context of message-passing neural networks remains under-explored. The research component of this project lies in the systematic study of whether, and to what extent, DP mechanisms can mitigate GNN-specific privacy risks. The core research questions include: 1. Can GNNs trained with differential privacy resist various attacks tailored for graph data? 2. Which types of protection are effective against particular attack vectors in GNNs, to what extent, and under what conditions do certain attacks remain successful despite these measures?

P02_PROJECT DETAILS	
Project Title	Python-Based Endpoint Detection and Response (EDR) System to Protect Target Systems from Simulated Attacks
Project Description	This project, supported by the Swinburne Cybersecurity Lab, aims to develop a lightweight and scalable Endpoint Detection and Response (EDR) solution for Linux-based virtual machines.The EDR system is intended to support the lab's cyberattack simulation and testing framework by detecting and responding to threats in a controlled environment.The solution will be built around OWASP Juice Shop and focus on five well-defined attack scenarios, including SQL injection, cross-site scripting, path traversal, broken authentication, and remote code execution. Each attack must be thoroughly documented, with all commands, tools, configurations, and indicators clearly recorded.For every attack, a corresponding defense must be implemented. These defenses will include specific detection rules and scripted automated responses, such as process termination or IP blocking. Defense procedures must also be fully documented with test evidence and validation logs.A modular Python-based EDR agent will integrate all attack and defense configurations. The agent will use structured configuration files to load detection rules and responses, and include monitoring components for files, processes, and network traffic. It will automatically trigger appropriate actions and maintain logs for evaluation.The final system will be validated through sequential testing of all scenarios, with performance metrics collected to assess effectiveness and inform any required adjustments. The result will be a reusable, documented EDR tool aligned with the lab's research and simulation needs.
Project Skills	☒ Cyber Security ☒ Programming (Software, Mobile and Web Development)
Environment	Kotlin, Python
Research Component	contemporary EDR systems and rule-based detection frameworks and how MITRE ATT&CK mappings have been leveraged in both academic studies and industry tools

P03_PROJECT DETAILS	
Project Title	Automated adversary emulation framework for cyber security exercises
Project Description	Develop a prototype of an automated adversary emulation framework. The aim of this project is to create a practical tool for functional (technical) cyber security exercise designers to efficiently generate accurate and customisable adversary Procedures, aligned with their exercise objectives and network architecture. The outputs of this tool will be used to inform the creation of forensic artefacts, commands, scripts and benign implants to enhance the credibility and realism of incident response exercises. The framework will be aligned to MITRE ATT&CK, and the team may reference existing industry standards such as MITRE Adversary Emulation Plans, Atomic Red Team, and Threat-based Purple Teaming when designing the prototype. The proof-of-concept will be demonstrated with APT40 Tactics, Techniques and Procedures. The prototype shall be designed to be flexible and scalable. It should be a viable long-term resource that allows users to generate Procedures for other threat actors, or adjust based on updated threat intelligence.
Project Skills	☒ AI/Machine Learning ☒ Cyber Security ☒ Programming (Software, Mobile and Web Development)
Environment	
Research Component	Existing adversary emulation frameworks and standards – e.g. MITRE Adversary Emulation and Red Teaming and Threat-Based Purple Teaming resources, other industry-recognised standards and frameworks.

P04_PROJECT DETAILS	
Project Title	Dark Web and Social Media Webpage Generator
Project Description	The project seeks to develop of recommendation system and suggestion dashboard. Students will be required to process the MITRE ATT&CK dataset to determine commonality and rareity in TTPs/Actors/Software to report on recommended exercises, build a item-based and collaborative or hybrid recommendation system. Student will be provided an exercise dataset to use along with the MITRE data.
Project Skills	☒ AI/Machine Learning ☒ Cyber Security ☒ Programming (Software, Mobile and Web Development)
Environment	Kotlin, Python
Research Component	Students will be expected to research various implementations of recommendation systems and the various data types supplied.

P05_PROJECT DETAILS	
P05_Project Title	Dark Web and Social Media Webpage Generator
Project Description	We are looking to create a small application that will allow for the generation (and customisation) of webpages to simulate cyber attacker infrastructure and other webpages. This project will be split into three components: research, template development, and dashboard development. Students are to research various attackers, including cybercriminals and state-sponsored cyber threat groups, and their infrastructure, including dark web forums and auctionrooms, via cyber advisories and reports, to discover real-life examples of webpages used in the field. Following this, students are to develop templated webpages (mainly HTML and CSS) inspired by this research. Finally, students are to create a GUI to allow for quick customisation of these templated webpages. If information is readily available and time permits, students may create a matrix of the most popular online forum(s), their purpose(s), and the threat actor(s) associated with it.
Project Skills	☒ Cyber Security ☒ Programming (Software, Mobile and Web Development)
Environment	☒ Microsoft (C#, C++, Windows, and database)
Research Component	Students will be expected to research a range of cyber security threat actors (including cyber criminals and state-sponsored groups) and their known infrastructure via cyber advisories and reporting – this will be to understand their current infrastructure, and also consider what may be used in the future.

P06_PROJECT DETAILS	
Project Title	AI-Integrated NDIS Provider Management Platform Leveraging Watsonx, IBM RPA, and Workflow Automation
Project Description	This project involves designing and prototyping an enterprise-grade NDIS (National Disability Insurance Scheme) Provider Management Platform that automates the full operational lifecycle using IBM Watsonx, RPA, and workflow automation. The solution will serve as a complete digital backbone for NDIS providers, covering everything from workforce acquisition to service delivery and compliance. Students will design and develop key modules for the following: 1. Job Posting and Talent Acquisition • Create job listings for support worker roles • Integrate with job boards or internal talent portals • Capture applications, resumes, and pre-screening forms • Build an AI-based ranking system for applicants using defined criteria 2. Staff and Participant Onboarding • Smart onboarding checklists for both staff and participants • Digital ID verification, form uploads, and contract generation • Workflow logic for assigning onboarding tasks to HR or coordinators • Integration with credential verification and police check providers 3. Shift Rostering and Real-Time Availability • Visual rostering calendar for coordinators • Support worker availability input via mobile/web • Conflict detection, shift overlap alerts, and travel time logic • Publish rosters and send notifications automatically 4. Compliance and Credential Tracking • Auto-alerts for expiring documents (e.g., First Aid, WWCC, NDIS Screening) • Real-time compliance dashboard for management • Restrict rostering of non-compliant workers • Integration-ready with third-party credential platforms 5. AI-Driven HR Ticketing and Support System • Integrated ticketing for HR and operations queries • Watsonx-powered Assistant (AskHR) to guide staff through policy or support • Categorize, prioritize, and auto-respond to common queries • Dashboard for HR staff to manage tickets and escalate where needed

	6. SIL (Supported Independent Living) House Management • Manage SIL homes, rooms, and resident assignments • Track shared shift coverage, house notes, and meal plans • Support documentation of routines, goals, and behavior support plans 7. Document Workflow and Incident Management • Build digital workflows for uploading, approving, and archiving service documents • Structured incident logging system for participants or workers (e.g., injury, behavioral) • Incident categorization, escalation paths, and notification logic • Attach supporting evidence, assign reviewers, and track resolution steps 8. Automated Communication (Email/SMS) • Trigger reminders for shifts, onboarding steps, document expiry • Notify participants/staff of changes, incidents, or critical alerts • Customizable templates for compliance and scheduling messages • Audit trail of all communications for reporting 9. Integration-Ready Architecture • Prepare system to integrate with Xero for payroll and invoicing • Build connector logic for NDIS portals or CareMaster (if applicable) • Use API-first approach for interoperability with government compliance systems Project Outcome: The project will result in a functional prototype with end-to-end NDIS workflow automation across HR, care delivery, scheduling, compliance, and support. Students will work in an agile environment with mentorship from Code24 and exposure to IBM tools (Watsonx, RPA, Orchestrate, Cloud Pak). The outcome will contribute to a commercial SaaS platform used by real-world NDIS providers in Australia.
Project Skills	AI/Machine learning; Cyber Security; Programming (software, mobile and web development)
Environment	IBM
Research Component	Research on AI in HR automation, shift scheduling, SIL workflows, compliance tracking, document workflows, API integration, and explainable AI for ethical automation.
Client support	Cloud (AWS, Azure, Google, Oracle) – Client must provide credits

P07_Project Title	Refer to P06: AI-Integrated NDIS Provider Management Platform Leveraging Watsonx, IBM RPA, and Workflow Automation
	Please note that Project 07 are looking for two groups.

P08_PROJECT DETAILS	
Project Title	AI-Driven Cybersecurity Automation
Project Description	Leveraging Machine Learning, Large Language Models (LLMs), Autonomous Agents, and Explainable AI for Proactive Threat Detection, Correlation, SOC Automation, and Incident Response Optimization In today's dynamic threat landscape, traditional cybersecurity methods are increasingly inadequate to detect and respond to sophisticated attacks. This project introduces an advanced framework combining machine learning (ML), large language models (LLMs), autonomous agents, and explainable AI (XAI) to build intelligent, transparent, and proactive cyber defense systems. 1. Proactive Threat Detection with Machine Learning Students will develop and train ML models that monitor and analyze high-volume telemetry from networks, endpoints, and cloud environments. Goals include: • Detecting anomalies and zero-day threats using unsupervised learning techniques. • Identifying known attack signatures (e.g., phishing, malware) via supervised models. • Continuously evolving detection logic using reinforcement learning. 2. LLMs for Threat Correlation and Triage Large language models will be utilized to streamline analysis of security data and open-source threat intelligence (OSINT). This component will enable: • Correlation of dispersed Indicators of Compromise (IoCs) into threat stories. • Natural language interpretation of alerts, logs, and threat reports. • Generation of incident summaries, playbook guides, and C-level reports. 3. Autonomous Agents for SOC Workflow Automation Students will simulate or implement intelligent agents that autonomously perform key SOC functions: • Investigating incidents by querying systems, analyzing logs, and cross-referencing threat data. • Triggering containment actions such as IP blocking or credential resets based on dynamic risk scores. • Auto-triaging tickets based on contextual severity and business risk. 4. Explainable AI for Decision Transparency and Trust To ensure transparency and accountability in security decisions, the project will integrate XAI methods such as: • SHAP or LIME to explain ML model predictions (e.g., why a connection was flagged as malicious). • Attention visualization in LLM outputs for analyzing token importance in decision-making. • Human-readable logic trees and audit trails for autonomous agent actions.

	This XAI layer will be essential for SOC analysts and auditors to validate and trust AI-driven outcomes. 5. Intelligent Incident Response Optimization By combining ML, LLMs, agents, and XAI, the system will optimize IR workflows through: • AI-assisted decision support for rapid containment and escalation. • Automated execution of consistent playbooks with full traceability. • Feedback-driven learning loops to continuously improve agent and model performance. 6. Real-World Use Case Simulations The project will implement and test real-world cybersecurity scenarios such as: • LLM-based natural language threat hunting across logs and endpoints. • Malware family clustering using unsupervised ML. • NLP-powered phishing detection on email and web content. • Dark web monitoring using LLMs and external intelligence feeds. • Explainable classification decisions for security audit trails. This project gives students hands-on experience in building cutting-edge cybersecurity tools, underpinned by industry-grade technologies like IBM Watsonx, Qradar, and open-source AI frameworks. With a focus on explainability, this approach ensures not only intelligent automation but also logical reasoning, transparency, and compliance — essential pillars of next-generation cybersecurity.
Project Skills	Using IBM Watsonx and Qradar tools AI and Machine Learning basics Understanding computer networks and security Detecting threats and suspicious activity Simple API and JSON data handling Writing simple scripts for security checks Python Codeing
Environment	Programming Language: Python AI Tools: IBM Watsonx Security Tools: IBM Qradar Platform: IBM Cloud Version Control: GitHub
Research Component	**Qradar** for event monitoring **Watsonx** for AI threat correlation **SOAR** tools for incident response automation

P09_PROJECT DETAILS	
P09_Project Title	**Accelerating Multi-user Language Models Watermarking schemes**
	The current AI era has witnessed an enormous daily volume of text and image generated by generative AI across the Internet. Such data is often distinguishable from human-authored content. However, detecting AI-generated content is important for various purposes, including intellectual property protection, ethical and social research, and ensuring genuine learning education. In July 2023, the White House issued an executive order to establish standards and best practices for detecting such AI-generated content. One promising approach is watermarking, which enables detection and traceability of generative AI outouts to support content authenticity. Although many existing works have been conducted, led by major industrial RnD efforts, such as Google's watermarking for the Lyria music generation model, and OpenAI's watermarking for content produced by Chat GPT and other statistical and applied-crypto based schemes, existing schemes face practical challenges of 1) supporting adaptive prompts from single or multi users interacting with language models, and 2) enabling efficient trackability to identify individuals of sources. A recent theoretical watermarking framework (Watermarking Language Models for Many Adaptive Users, https://eprint.iacr.org/2024/759.pdf, IEEE S&P'25) has formalised the above challenges. However, the framework's performance is currently limited at watermarking and tracing on a text block by block, without considering the semantic relationship between text blocks within a single user interaction session. Furthermore, there is no existing implementation of this design. This project aims to be the first to evaluate the real-world performance of existing LLM watering schemes on different open-source language models using public image and text datasets. Specifically, the project will 1) Study existing watermarking schemes for LLMs, and assess their properties w.r.t. soundness, completeness, robustness, and undetectability. 2) Re-implement the above watermarking framework on Swinburne's HPC platform 3) Improve watermarking and tracing operations by applying cryptographic techniques and efficient data structures, in both single-user and multi-user settings.
References	Cohen et al. Watermarking Language Models for Many Adaptive Users, IEEE S&P'25 Gligoroski et al., An LLM Framework For Cryptography Over Chat Channels, https://arxiv.org/abs/2504.08871 Apr 2025

P10_PROJECT DETAILS	
P10_Project Title	**Data Deduplication in Machine Unlearning**
Project Description	Data deduplication is a process of identifying and cleaning up duplicated dataset during the training of machine learning models. This process not only optimises the computational complexity but also reduces the risk to data privacy. Therefore, duplication has been widely adapted in federated learning and reinforcement learning settings. Under the European General Data Protection Regulation (GDPR), end-users have the right to request the removal of their personal data from a trained model if their data was used during the training process. While existing studies have proposed exact or approximate methods for unlearning revoked data and verifying the effectiveness of unlearning [1,2], the impact of data duplication on the unlearning process, especially from an adversarial perspective, remains largely unexplored [3]. In particular, an adversary could duplicate a subset of the training data used in the target model and later incorporate it again into the training dataset. If the attacker subsequently requests the model owner to unlearn this duplicated dataset, it may still degrade the model's performance. However, the extent of adversarial effort required (e.g., number of adversarial clients, amount of training data, and time needed) to cause a significant impact has not yet been experimentally studied [5]. This project aims to systematically investigate the security implications of data deduplication in a federated learning setting. The objectives include: 1. **Evaluation of Existing Methods:** Replicate and evaluate the implementation of existing works [2,3] on various deep learning architectures (e.g., LeNet, AlexNet). 2. **Adversarial Parameter Analysis:** Explore the effects of different adversarial configurations, including the number of malicious clients, training data volume, and training duration. 3. **Heterogeneous Data Settings:** Adapt experiments to federated learning environments with non-IID (non-identically and independently distributed) data to reflect realistic scenarios. 4. **Defense Evaluation:** Assess the effectiveness of data duplication attacks against current federated learning robustness defenses. This includes incorporating theoretical frameworks and existing defense strategies such as FLShield [4]. 5. **Recommendation Framework:** Develop a comprehensive defense recommendation framework to mitigate the risk of data deduplication attacks in federated learning systems.
Research Component	[1] Xu et al. Machine Unlearning: A Survey, ACM Computing Survey 2023, https://dl.acm.org/doi/10.1145/3603620

	[2] Cao et al. FedRecover: Recovering from Poisoning Attacks in Federated Learning using Historical Information, IEEE S&P'23, https://arxiv.org/abs/2210.10936 [3] Ye et al. Data Duplication: A Novel Multi-Purpose Paradign in Machine Unlearning, Usenix Security 2025, https://www.usenix.org/system/files/conference/usenixsecurity25/sec25cycle1-prepub-783-ye-duplication.pdf. Github. https://zenodo.org/records/14736535 [4] Kabir et al. FLShield: A Validation Based Federated Learning Framework to Defend Against Poisoning Attacks, IEEE S&P'24. [5] Shejiwalkar, Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning, IEEE S&P'22.

P11_PROJECT DETAILS	
Project Title	**Reinforcement Unlearning Attacks and Defence**
Project Description	Reinforcement learning (RL) is a practical agent-based learning paradigm that trains agents to make optimal decisions within a given environment over time using reward functions. This approach is widely adopted in AI systems and robotic agents. Under the European General Data Protection Regulation (GDPR), end-users have the right to request the removal of their personal data from a trained model if it was used during the training process. However, the issue of *unlearning attacks* in agent-based reinforcement learning has remained largely unexplored in recent years. Specifically, end-users should have the ability to request that a trained agent "forgets" private data samples that were previously used during training. Recent work has proposed unlearning schemes that aim to preserve agent performance in simulated environments [1]. These designs typically rely on decremental reinforcement learning during the agent training process and include *environment poisoning attacks* as part of the strategy. Furthermore, the concept of *environment inference attacks* has been introduced to assess whether the unlearning was successful. This concept uses genetic algorithm-based attacks to evaluate whether the agent has effectively erased knowledge of the removed environment [2]. However, the computational overhead and memory requirements of these approaches have not yet been thoroughly investigated. Therefore, this project aims to: 1. **Evaluate existing benchmarks** of proposed unlearning schemes in practical simulation environments. 2. **Develop an assessment framework** to test these schemes on real-world robotic hardware, to determine their readiness for real-world deployment.

	To achieve these goals, the project will follow the step-by-step methodology outlined below. 1. **Evaluation of Existing Methods:** Replicate and evaluate the implementation of existing works [1]. 2. Conduct environment inference attacks using generic algorithms [1,2]. 3. Conduct the implementation on different agent simulation environments, including RLBench (Imperial College London), MetaWorld (Stanford, UC Berkeley, Google), VirtualHome (MIT, Toronto) for LLM-integrated task planning, and RH20T (SJTU) with environment inference metrics. 4. Empirically measure the computational time, network latency, and memory overhead of this design in practical robotic hardware specifications like Kinova Robotic and UR5e robotic hardware. 5. Benchmark the practicality of these scheme design against real-world constrained requirements of robotics: limited memory, and network throughputs, and the frequency of unlearning in practice. **6.** Provide a recommendation framework in practice for further unlearning reinforcement design.
Research Component	[1] Ye et al. Reinforcement Unlearning. In NDSS'25. https://www.ndss-symposium.org/wp-content/uploads/2025-80-paper.pdf, Github. https://github.com/cp-lab-uts/Reinforcement-Unlearning [2] Pan et al. How You Act Tells a Lot: Privacy-Leaking Attack on Deep Reinforcement Learning, AAMAS 2019.

	P12_PROJECT DETAILS
Project Title	**Privacy risk estimation using AI approach in encrypted database** **cy risk estimation using AI approach in encrypted database**
Project Description	Searchable Encryption with encrypted (cloud) databases enables end-users to issue encrypted queries and still retrieve matching records without revealing the query content to the untrusted server hosting the encrypted database [1]. This approach has recently been adopted in platforms such as MongoDB [1] and the AWS Database Encryption SDK [2]. However, such designs remain vulnerable to *leakage-abuse attacks*, in which attackers exploit patterns such as the frequency of repeated encrypted queries and corresponding results—combined with auxiliary knowledge—to infer private information [3,4]. Existing approaches to detect and prevent leakage-abuse attacks typically require comprehensive frequency analysis over large datasets, making them unsuitable for real-time risk assessment. To address this, a recent framework called **ALERT** [5], along with its implementation [6], has been proposed. ALERT

	is a machine learning-assisted tool that provides on-the-fly, quantitative estimates of privacy leakage risk. Therefore, this project aims to: 1. **Evaluate Existing Methods**: Replicate and evaluate the implementation of ALERT [5] on Swinburne's High Performance Computing (HPC) infrastructure. 2. **Dataset Availability Testing**: Adjust the proportion of training data available to assess the robustness and practicality of ALERT under realistic data scarcity scenarios. 3. **Streaming Database Adaptation**: Design a simplified extension of ALERT to support *dynamic streaming databases*, where data arrives in batches rather than as a static dataset. 4. **Framework Enhancement**: Investigate whether continual learning and explainable AI frameworks, such as SHAP, can enhance ALERT's performance for real-time learning and interpretation.
References	[1] Queryable Encryption, MongoDB, https://www.mongodb.com/docs/manual/core/queryable-encryption/ [2] Amazon encryption SDK, https://docs.aws.amazon.com/database-encryption-sdk/latest/devguide/searchableencryption.html [3] Cash et al. Leakage-abuse attacks against searchable encryption, ACM CCS' 2015. https://dl.acm.org/doi/10.1145/2810103.2813700 [4] Falzon et al., Full database reconstruction in two dimensions. In ACM CCS' 2020. https://dl.acm.org/doi/10.1145/3372297.3417275 [5] Wang et al. , ALERT: Machine Learning-Enhanced Risk Estimation for Databases Supporting Encrypted Queries, Usenix Security 2025. https://www.usenix.org/system/files/conference/usenixsecurity25/sec25cycle1-prepub-190-wang-longxiang.pdf [6] https://doi.org/10.5281/zenodo.14726862