Plan Overview

A Data Management Plan created using DMPonline

Title: Agentic AI-Enabled penetration testing for SMEs.

Creator: Faisal Aldaij

Principal Investigator: Faisal Aldaij

Data Manager: Faisal Aldaij

Project Administrator: Faisal Aldaij

Affiliation: University of Plymouth

Template: DCC Template

Project abstract:

The PhD study focuses on automating penetration testing for SMEs by utilising Artificial Intelligence (AI), where an AI-based system will be designed to autonomously emulate real-world adversary behaviour across all stages of the cyber kill chain. In this context, Agentic Artificial Intelligence refers to AI-based systems that autonomously operate with goal-oriented capabilities, capable of making decisions, and adapting to the surrounding environment. Accordingly, Agentic AI-Enabled penetration testing aims to address the evolving threat landscape and advanced cyberattacks, particularly for SMEs that lack specialised security teams or infrastructure, as the ultimate goal.

ID: 177565

Start date: 01-04-2025

End date: 01-04-2029

Last modified: 01-08-2025

Agentic AI-Enabled penetration testing for SMEs.

Data Collection

What data will you collect or create?

I will collect and create data as below:

1- Synthetic data will be generated using simulated penetration testing scenarios with tools like Metasploit and BurbSuit.

2- Agents' logs and interactions with the system.

3- Large Language Models workflows scripts, linking codes, process of the simulated steps of the attacks.

4- Setup files and system environment settings.

5- Documented results of the simulated attack scenarios.

6- Spreadsheets gathering the output of the literature review including the review process and criteria as well as the analysis.

How will the data be collected or created?

The data will be created through:
1- Simulated Cyber environment using virtual machines and Virtual Boxes.
2- Systematic selection of published academic papers and resources to build understanding of the research topic.
3- Logging files embedded within the experimental system to record the agent interactions and behaviour.

Documentation and Metadata

What documentation and metadata will accompany the data?

1- README files describing each dataset (content, and usage).

2- Metadate for each file shows (Creation date, tools used, and versions)

3- Experiment setups details will be stored in (JSON) format for configuration management.

4- Git version control logs that document how the source code has evolved over time.

Ethics and Legal Compliance

How will you manage any ethical issues?

Only simulated system and non-sensitive data will be used.
All experiments will be carried out in a sandbox (Controlled virtual environment)
The research will adhere to the university’s ethical principles framework
No identifiable information will be used in the research.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

1- Copyrighted materials that will be used in this research, will be clearly cited and used under fair use guidelines.
2- All tools, libraries, and datasets that will be used in the research will be open-source and cited appropriately.
3- The research output (code and datasets) will be owned by the researcher and the University of Plymouth within open-source distribution (MIT) when applicable.

Storage and Backup

How will the data be stored and backed up during the research?

1. The university cloud storages (OneDrive, and SharePoint) will be utilized to store the research data.
2. Scheduled Local backup will be maintened on both enternal and external encrypted hard drive.

How will you manage access and security?

1. Multi-Factor authentication will be enabled on all cloud accounts.
2. Access to the research data and codes will be limited to authorized personnel (Researcher, and Supervisors) with password protection.
3. Encryption containers will also be implemented for any sensitive information.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

1- The redult of the experiments, Agent Models, and Synthetic datasets.
2- Codes and Scripts.
3- AI Agents logs shows the agents behaviour and interaction with the systems.
4-The experiment documentations and metadata to enable the resuse and further improvement.

What is the long-term preservation plan for the dataset?

1- PEARL the university repository will be used.
2- JSON, PDF, and TXT formats will be used for long term use.
3- To ensure appropriate citation and further resuse, presistent identifiers (DOI) will be assigned.

Data Sharing

How will you share the data?

1- The university and GitHub repository will be utilized to make the datasets, codes, and documentation publicly available.
2- Academic publication and open access conferences will be used to share the research results.

Are any restrictions on data sharing required?

1- Ethical and legal guidelines will be considered when sharing.
2- To avoid possible misuse, sensitive material may be shared only with verified researchers upon official request.

Responsibilities and Resources

Who will be responsible for data management?

1- (Myself), I will be the primary responsible for DM with direction from my supervisor.
2- Guidance and advices may request from the university RDM when needed.

What resources will you require to deliver your plan?

1- Cloud and backup services provided by the university.
2- Hardware (Computer, internet access) for experiment including (training the model, and environment simulation).
3- Access to academic databases for literature review purpose.
4- Software Licences.
5- GitHub for tracing the code evolvement and collaboration.