# SLDP Intern Action Learning Project Brief

## Project 3: AI Agent Guardrails Service

**Project Title:**  
a Guardrail Service for Responsible AI Agent Use

**Key Stakeholders:**

|  |  |  |
| --- | --- | --- |
| **Originating/Hosting Department** | **Project Sponsor** | **Main Contact Person** |
| iDNA | Dion Dodgen TBC | Denzil Jacobs (iDNA) |

**Project Brief:**

* **Context:**  
  AI agents must operate within ethical and secure boundaries, especially in sensitive banking environments. Guardrails help ensure safe, compliant, and responsible AI behaviour.
* **Problem Statement:**  
  Design a guardrail service that monitors and filters AI agent **inputs OR\* outputs**. The service should detect and mitigate risks such as data leakage, bias, inappropriate content, jailbreaking or malicious requests.

*\* Teams may choose to focus on specific scenarios.*

* **Outcome or Question:**  
  Deliver a working prototype of a guardrail service that can be integrated with AI agents. Include rule-based and/or ML-based and/or GenAI-based safeguards, and demonstrate its effectiveness through testing. All scenarios should be accompanied by possible remediation paths (i.e what happens if an input/output is flagged by the guardrail service).

**Expected Final Deliverables (Apart from final presentation):**

* Guardrail service prototype (codebase and documentation)
* Test results and evaluation report
* Demo video or live demonstration