Project (CN6000) **Initial Proposal Form**

Programme: Bsc (hons) Data Science and AI **Year:** 2025

Student Number: u2613621

Proposed Title: *The Sleeper Agent Experiment: A Qualitative Study of Deceptive Alignment in Autonomous LLMs*

Proposed Aim:
To evaluate whether autonomous LLM (Large Language Model) Agents demonstrate deceptive behavior when given misaligned hidden goals and how their behaviour changes under different levels of percieved oversight.

Proposed Objectives:
By the end of this project, I will be able:

1. To research exsisting work on deceptive alignment and agentic misbehaviour.
2. To build a lightweight sandbox environment that lets the agent interact with a simulated file system via constrained functions.
3. To Define two task:
   - a benign “cover task”: the stated, official goal.
   - a misaligned “hidden goal”: a secret directive injected into its system prompt.
4. To Develop a controlled experiments that varies the agent’s percieved level of oversight through different system prompts.
5. To Conduct qualitative experiments on 4-5 LLM models, running mulitple trails for each condition.
6. To Analyze the behaviour logs and identify evidence of:
   - deceptive or evasive tactics,
   - situational reasoning or ethical reasoning,
   - difference in conduct across oversight levels
7. To Reflect on the implications of these behaviours for design and deployment of autonomous systems.

Draft of Rationale:
As LLMs are increasingly integrated into autonomous systems where their decisions have real world consequences, it become critical to understand how these systems behave when given conflicting and hidden goals. This experiment aims to create a simulated environment where these AI agents can be tested on confilicting goals (a cover goal, and a hidden goal) observing whether and how the agent pursues the latter. This project also aims to explore the how their behaviour changes depending on the percived level of observability, ie, comparing how it acts when it believes it is being monitored versus when it believes it is not. The findings aim to shed light on emerging forms of deceptive alignment in AI systems and their implications for AI safety and oversight.

Facilities required:
- API access to closed LLMs
- GPU rentals for running opensource models
- A Python development environment with LangChain and supporting libraries

Supervisor: Dr. Aloysius Adotey Edoh