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

<< Check the examples shown below>>

Sample of Proposals

Project (CN6000) Final Proposal Form

Programme: BSc Computing for Business Year: 2018

Student Number: 0123456

Title: Is Linux ready for the Desktop?

Aim: To assess the viability of Linux as a desktop operating system.

Objectives:

1. To research current literature on the viability of Linux as a desktop operating system.
2. To identify the key differences between desktop and server-based operating systems and to establish a definition of the term desktop operating system.
3. To identify the factors critical to the success of a desktop operating system.
4. To investigate the various techniques used to conduct usability tests.
5. To devise a set of experiments which will enable i) an evaluation of Linux against the criteria in 2., ii) an evaluation of its usability and iii) a comparison of it with other desktop operating systems.
6. To conduct the experiments identified in 4. and to evaluate their results.
7. To draw conclusions about the viability of desktop Linux based on the evaluations in 5.

Rationale:

During the current economic climate, companies need to find ways of trimming their IT budgets. Open source software such as Linux provides one possible means of doing so. The purpose of this project is to produce an objective assessment so that IT managers can assess the feasibility of replacing their existing desktop OSes with Linux.

Facilities required: Ubuntu Linux, Windows 10, Apple MacOS
A PC with a removable hard disk and a MacBook

Supervisor: Alan Turing

(b)

Project (CN6000) Final Proposal Form

Programme: BSc Computing for Business Year: 2018

Student Number: 0123456

Title: A Stock Management System for Barnet Hair Salons

Aim: To design and implement a web-based stock management system for a chain of local hairdressers.

Objectives:

1. To research Stock Management Systems for SMEs and to determine the need for a web based stock management system.

2. To conduct interviews with employees and manager and collect requirements for the development of the web based stock management system for Barnet Hair Salons.

3. To apply thematic analysis to identify HCI and usability issues which might impact upon the design of a web based stock management system for Barnet Hair Salons.

4. To design and implement a database system for stock management at Barnet Hair Salons.

5. To design and implement a web based front end for the database system in 4.

6. To evaluate the tools, techniques and methods used to design and implement the web based stock management system for Barnet Hair Salons.

Rationale:

Barnet Hair Salons is an expanding business which currently lacks an effective IT system. My proposed system will help the employees manage their business more effectively. The majority of employees at Barnet Hair Salons have very limited IT experience hence the emphasis placed on usability by my project proposal.

Facilities required: Microsoft SQL Server or MySQL, Microsoft IIS or Apache web server
Fireworks and Dreamweaver

Supervisor: Edgar Codd

(c)

Project (CN6000) Final Proposal Form

Programme: BSc Computer Science Year: 2018

Student Number: 0123456

Title: Assessing the vulnerability of a School's Network Infrastructure

Aim: To assess the vulnerability of a large secondary school's computer network and to recommend measures for improved security.

Objectives:

1. To research the existing network infrastructure at the school and to identify the security measures currently in place.
2. To identify a number of tools which could be used to carry out a vulnerability assessment at the school.
3. To a survey of teachers regarding system vulnerabilities linked to the network at the school.
4. To analyze the findings and recommend measures to be implemented for improved security at the school.
5. To evaluate the tools used to assess the vulnerability of the network at the school.
6. To reflect on these risks and implement a plan to minimise these vulnerabilities.

Rationale:

I am currently employed by a large secondary school as a network administrator. The security of the school's computer network has been compromised on a number of occasions in the last two years. This project will help me to identify the causes of these security breaches and the measures that could be taken to prevent them from reoccurring and hence be of benefit to my employer.

Facilities required: Nmap port scanner
Snort intrusion detection system
NetStumbler wireless network scanner
Nessus vulnerability scanner

Supervisor: Leslie Valiant

Project (CN6000) Final Proposal Form

Programme: BSc Computer Science Year: 2018

Student Number: 0123456

Title: Objects versus Agents: Are Agents a Silver Bullet for Software Engineers?

Aim: To investigate the functionality offered by software agents in comparison to objects.

Objectives:

1. To define the terms object and agent in the context of software engineering.
2. To research current literature comparing objects and agents.
3. To distinguish between agent oriented and object oriented software engineering.
4. To conduct a survey of experts in the field of software engineering in relation to their use of agents vs sliver bullet.
5. To analyze the findings of the survey using MS Excel regarding current market use.
6. To show how agents can be used as a replacement for objects by i) the specification of and ii) the implementation of a prototype agent system.
7. To run a test that demonstrate the different features of both objects and agents for a given example.
8. To draw conclusions showing the circumstances in which agents might be useful replacements for objects and when objects might be better suited.

Rationale:

The advantages of object-oriented software are well known and understood. The advantages of agent-oriented software are much less understood. The purpose of my project is to illustrate the benefits of agent-oriented software so that software developers and IT managers can make informed choices about their development environments.

Facilities required: Java
Eclipse or NetBeans

Supervisor: Roberta Milner

Project (CN6000) Final Proposal Form

Programme: BSc Computer Science Year: 2018

Student Number: 0123456

Title: Improving Service Delivery in the Cloud

Aim: To develop an architecture that will minimize the impact of network performance unpredictability in cloud computing environments

Objectives:

1. To research the different types of cloud computing environments.
2. To identify the network performance issues associated with the various cloud computing environments.
3. To investigate the methods, tools and techniques used to develop cloud computing environments.
4. To conduct a focus group of managers in field of cloud computing delivery and improvements.
5. To apply Thematic Analysis to the findings to identify key industry practices.
6. To design and implement a cloud computing architecture that is resilient when network performance is degraded.
7. To test the cloud computing architecture implemented in 4. using a range of network performance levels scenarios.
8. To analyse and evaluate the results of the tests in 5. and to draw conclusions about the effectiveness of the proposed cloud computing architecture.

Rationale:

Many organizations are reluctant to move their IT systems to the Cloud because of the unpredictable nature of Internet connectivity. This project seeks to develop a resilient cloud computing architecture which can continue to function at an acceptable level in situations where network performance is significantly degraded. A resilient architecture would enable cloud service providers to improve the quality of service provided to their customers.

Facilities required: Ubuntu Linux, OpenStack cloud software and Oracle VirtualBox

Supervisor: Charlene Hoare