From Prediction to Action: Personalized Explainable AI Profiles and Robustness Testing for At-Risk Students in Online Education

Supervisor: Mr. Kevin Ryan
Programme: MSc in Data Science
Email: sara.iqbal@city.ac.uk
Student Number: 240053369

Introduction

Online learning of structures plays a vital role in contemporary-day training by presenting flexible and scalable get right of entry to to gaining knowledge of. However, a significant challenge in online training is the high rate of pupil dropout, which is regularly related to the dearth of well-timed and customized support. To cope with this issue, many establishments are the usage of device gaining knowledge of (ML) models to predict which students are susceptible to dropping out. While those models are regularly accurate, they tend to paintings like "black boxes," supplying little explanation of their predictions. As a result, educators are left without the ability to make knowledgeable and centered interventions.

This dissertation aims to solve this problem by developing personalised explainable AI (XAI) profiles. These profiles will not only predict the risk of student dropout but will also explain why a particular prediction was made. The project will use the Open University Learning Analytics Dataset (OULAD), which contains anonymised student information on demographics, academic performance, and engagement with the virtual learning environment (VLE). Two key explainability tools, SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), will be applied before and after model training to provide both general and individual-level explanations of the model's behaviour.

In addition to improving interpretability, the project will apply unsupervised clustering techniques, such as K-Means, to group students with similar characteristics based on their learning behaviour and demographics. This will help personalise the explanations further by linking them to student groups or "personas"—for example, a group of students who perform well but show low engagement. To ensure the reliability of the explanations, the project will also include robustness testing. This involves making small changes to input data (such as increasing the number of clicks) and checking if the explanations remain consistent.

The broader framework for this dissertation brings together ideas from computer science, statistics, and education. As shown in Figure 1, the methodology lies at the intersection of these three fields, combining machine learning, deep learning, educational data mining, and statistical analysis. Explainable AI serves as the core technique that links these disciplines, to make predictive models not only accurate but also transparent and useful for real-world educational decision-making.

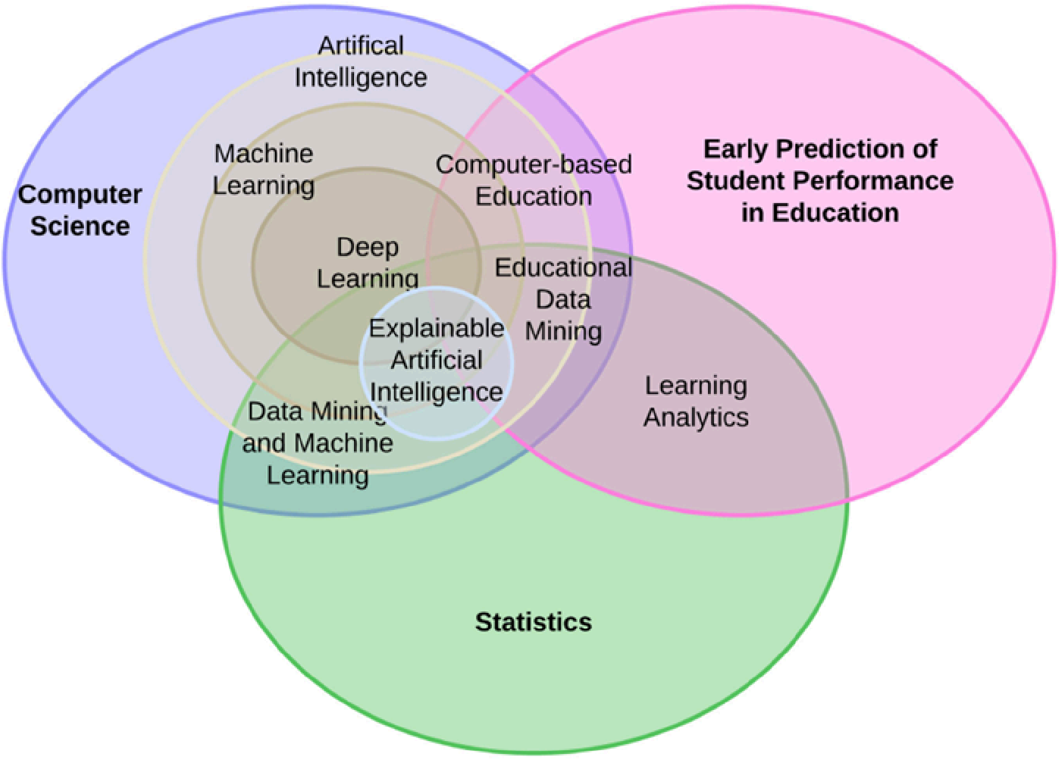

Figure 1. *Interdisciplinary Foundations of Explainable AI in Educational Prediction ([2024]).* This Venn diagram demonstrates the convergence of AI, statistics, and educational analytics in early student performance prediction.

The final product will be a system that generates **individualised risk profiles** for students, offering both predictions and their justifications. These profiles will help educators better understand at-risk students and tailor interventions accordingly. By making AI decisions more transparent and trustworthy, this dissertation seeks to bridge the gap between predictive modelling and actionable educational support—making a meaningful impact in online learning environments.

Research Question

1. How accurately can machine learning models predict student dropout in an online learning environment using the OULAD dataset, and which features most strongly influence these predictions?

2. To what extent do SHAP and LIME provide consistent and interpretable explanations for both baseline and advanced models, and how do these explanations vary across behavioural clusters of students?

3. How robust are the model explanations produced by SHAP and LIME when subjected to minor perturbations in input features, and what does this imply about their reliability in educational contexts?

Objectives

- To develop and evaluate machine learning models for identifying students at risk of dropping out, utilising the Open University Learning Analytics Dataset (OULAD).

- To apply Explainable AI techniques, specifically SHAP and LIME, in order to interpret model decisions both prior to and following training.

- To implement clustering algorithms to segment students based on demographic and behavioural features, thereby enhancing the personalisation of model interpretations.

- To assess the robustness and stability of model explanations through adversarial testing by introducing minor perturbations to input features.

- To generate individualised risk profiles for students that integrate predictive outcomes and interpretability insights, supporting educators in delivering targeted interventions.

- To ensure that the proposed methodology remains interpretable, ethically sound, and scalable for real-world deployment in educational settings.

Critical Context

The continued growth of online training has introduced both opportunities and challenges, specifically regarding student engagement and retention. One of the maximum urgent challenges is the excessive dropout rate amongst online learners. Educational institutions are increasingly turning to machine learning (ML) to identify early warning signs and are expecting dropout risks. However, as Choi et al. (2023) point out, the lack of transparency in lots of ML models hinders their adoption through academic stakeholders who require interpretability to guide significant intervention.

To cope with this issue, the latest studies have focused on Explainable Artificial Intelligence (XAI), which seeks to offer transparency into the decision-making techniques of ML systems. Techniques consisting of SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) have turned out to be outstanding for providing global and local model interpretability. Gunasekara and Saarela (2025) verified using SHAP to produce student-level explanations of dropout risk, enabling educators to understand why positive college students are categorised as at-risk. Similarly, the take a look at “On the Use of Explainable Artificial Intelligence to Evaluate School Dropout” (2022) proposed quantitative metrics to evaluate the readability and equity of XAI motives, reinforcing the need for interpretability in real-world applications.

Complementary to explainability, using hybrid ML models has been verified to be powerful for balancing accuracy and transparency. The take a look at titled A Novel AI-Driven Model for Student Dropout Risk with Explainable AI Insights (2024) provided a pipeline that consists of conventional classifiers (e.g., logistic regression), neural networks, and a hybrid logistic regression-ANN–ANN (HLRNN) model. SHAP is then used to interpret model outcomes. This shape demonstrates a realistic and interpretable framework for dropout prediction and has stimulated the methodological basis of this dissertation.

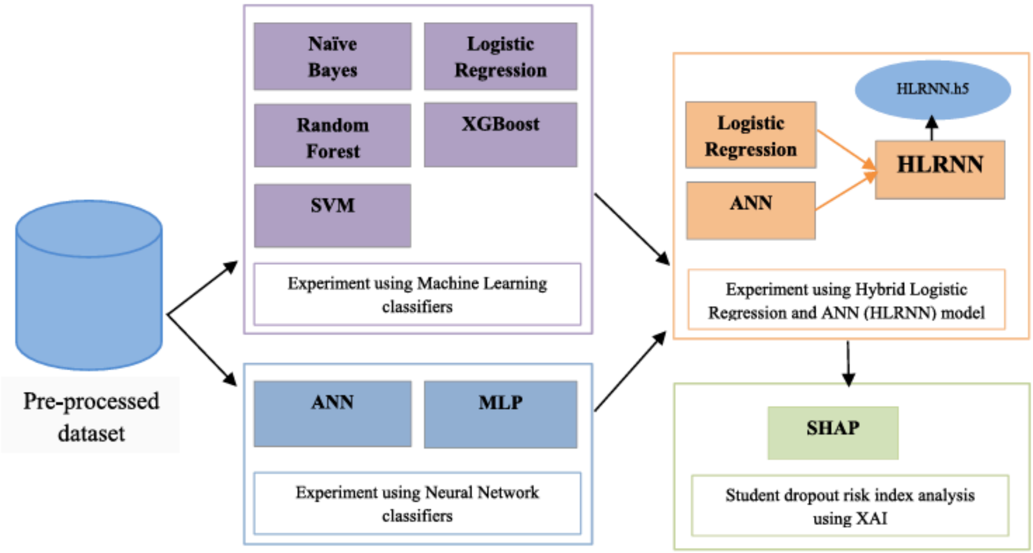
Naïve Bayes
Logistic Regression
Random Forest
XGBoost
SVM
Experiment using Machine Learning classifiers
Pre-processed dataset
ANN
MLP
Experiment using Neural Network classifiers
HLRNN.h5
Logistic Regression
HLRNN
ANN
Experiment using Hybrid Logistic Regression and ANN (HLRNN) model
SHAP
Student dropout risk index analysis using XAI

Figure 2: Workflow adapted from “A Novel AI-Driven Model for Student Dropout Risk with Explainable AI Insights” (2024), illustrating a multi-model approach combining machine learning, neural networks, and SHAP explanations for dropout risk analysis.

Figure 2 offers a visual representation of a layered ML and XAI architecture that parallels this dissertation’s design. Like the referenced study, this research begins with basic classifiers and evolves toward more complex models, incorporating SHAP for model interpretation. However, this dissertation builds upon that structure by incorporating both SHAP and LIME at pre- and post-training stages, offering a more comprehensive evaluation of feature importance.

Clustering provides some other crucial measurements to personalising factors. The Interpret3C framework (2024) delivered a clustering approach the uses of customized function selection, which complements the contextual relevance of student segmentation. This aligns with the modern-day study`s use of K-Means clustering to generate behavioural profiles (e.g., low-engagement or high-overall performance groups), permitting educators to interpret version outputs within the context of comparable scholar behaviours.

Finally, current literature identifies a great hole in the robustness evaluation of XAI factors. While SHAP and LIME offer interpretability, their factors may be volatile below minor modifications to the input features. This project addresses that issue through incorporating robustness testing. As stated in current XAI studies, strong factors need to stay constant below small, managed enter perturbations to be deemed straightforward for decision-making.

In summary, this dissertation synthesises insights from at least five key research contributions: (1) predictive modelling in education (Choi et al., 2023), (2) SHAP-based interpretability (Gunasekara & Saarela, 2025), (3) explainability metrics (2022), (4) hybrid ML pipelines (2024), and (5) clustering-based personalisation (Interpret3C, 2024). By combining these approaches into a cohesive framework that includes prediction, explanation, personalisation, and robustness, this research contributes a novel, practical, and interpretable solution for addressing dropout risks in online education.

Approaches: Methods & Tools for Design, Analysis & Evaluation

This research adopts a mixed-strategies technique that mixes supervised machine learning, unsupervised clustering, and explainability strategies to predict and interpret scholar dropout risk. The examine uses the Open University Learning Analytics Dataset (OULAD), which includes anonymised facts such as demographics, evaluation overall performance, and digital learning environment (VLE) interactions.

Data Collection and Preprocessing:
The project begins through obtaining and cleansing the OULAD dataset. Preprocessing consists of addressing missing values, outlier detection, and standardisation of numeric capabilities, along with the quantity of VLE clicks and delays in evaluation submissions. Categorical capabilities like gender, region, and disability status may be one-hot encoded. Additionally, a binary target variable will be derived from the final_result column to categorise students into dropout and non-dropout categories. This step guarantees the dataset is version-prepared and that the capabilities are appropriate for each education and interpretability.

Baseline Model and Pre-Training Explanations:
A logistic regression version will be used as a baseline for early-level rationalization and overall performance benchmarking. This version, at the same time as simple, permits for interpretable coefficients and acts as a basis to understand characteristic relationships. SHAP and LIME may be implemented in this version to discover which capabilities make contributions to predictions, previous to education greater complicated educational models. This establishes a preliminary layer of transparency and permits for comparisons with post-education interpretability.

Training Advanced Models:
Following the baseline, more powerful machine learning models, along with Random Forest and XGBoost may be trained on the processed dataset. These models are selected for his or her capacity to seize non-linear relationships and interactions amongst capabilities. Model overall performance will be evaluated the usage of a couple of metrics which including accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices, to offer a complete knowledge of predictive quality.

Post-Training SHAP and LIME Analysis:
Once the models are trained, SHAP and LIME will again be used to interpret predictions. SHAP will provide both global feature importance (how features contribute overall to model output) and local explanations (how features influence individual predictions). LIME will be used for case-specific explanations by perturbing input features and observing their impact on predictions. This dual application allows for validating model transparency and consistency in feature influence before and after model refinement.

Clustering for Behavioural Segmentation:
To personalise interpretation, K-Means clustering will be employed to segment students based on behavioural patterns (e.g., VLE activity levels) and demographic traits. This unsupervised learning step generates student personas—such as high-performing but low-engaged learners—which contextualise SHAP and LIME outputs. For instance, the dropout prediction of a low-engagement student can be interpreted through the lens of that behavioural cluster, making the insights more actionable for educators.

Comparative Analysis of Explanations:
A key novelty in this study is comparing the feature importance outputs from SHAP and LIME before and after model training, as well as across different clusters. This comparison will reveal whether certain features consistently impact predictions, or if the model's logic shifts across different student groups. Such insight contributes to building trust in the model and understanding its fairness.

Robustness Testing Using Adversarial Perturbation:
To evaluate the stability of model explanations, adversarial testing will be conducted. This involves making minor, controlled changes to input features—such as altering the number of clicks or shifting the age band—and then monitoring the resulting explanations from SHAP and LIME. A robust model should yield stable explanations for small perturbations. If not, it indicates interpretability fragility, which will be documented as a risk.

Generating Personalised Risk Profiles:
The final goal is to develop a personalised risk profile for each student that integrates prediction outcomes, clustering results, and SHAP/LIME explanations. These profiles will summarise why a student is predicted to be at risk, highlighting key contributing features and situating them within their cluster. This practical output enables educators to understand not only which students are at risk, but also the reasons behind those risks in a meaningful, personalised context.

Tools and Implementation:
All analyses will be implemented using Python, with libraries including pandas, scikit-learn, xgboost, shap, and lime. Experiments will be conducted using Jupyter Notebooks on Google Colab. If feasible, a prototype dashboard may be developed using Streamlit or Dash to demonstrate how educators could interact with the personalised insights.

Ethical, Legal, and Professional Considerations:
The OULAD dataset is fully anonymised and publicly accessible, mitigating direct ethical concerns. As no personally identifiable data will be collected or processed, the research will require only Part A of the City University Ethics Review Form. The project complies with

principles of data protection, ethical AI, and fair use, safeguarding all users' emotional, intellectual, and privacy-related well-being.

This comprehensive, step-by-step approach ensures the research is both technically rigorous and practically applicable, contributing to the development of transparent, interpretable, and personalised AI tools for educational support.

Work Plan

The work plan spans an established 16-week timeline, starting off on 9th June and concluding with submission by 1st October. It (Figure 3) outlines the most important challenge phases, aligned with dissertation milestones and academic deliverables. Tasks are logically sequenced to house dependencies (e.g., Data Preprocessing, then creating a baseline of ML models with pre-training of SHAP/LIME followed by advanced training of MI model and comparing them with Post Traing SHAP/LIME then while performing clustering it will help me to create a personalized risk profiles of student), and a three-week buffer is included to manage unexpected delays and refine the very last output.

The Gantt chart illustrates all key ranges from literature evaluation to chance profiling, robustness testing, and the very last submission. The plan guarantees every student's goal is addressed in a well-timed and methodical manner, at the same time as also permitting room for iterative improvement and evaluation

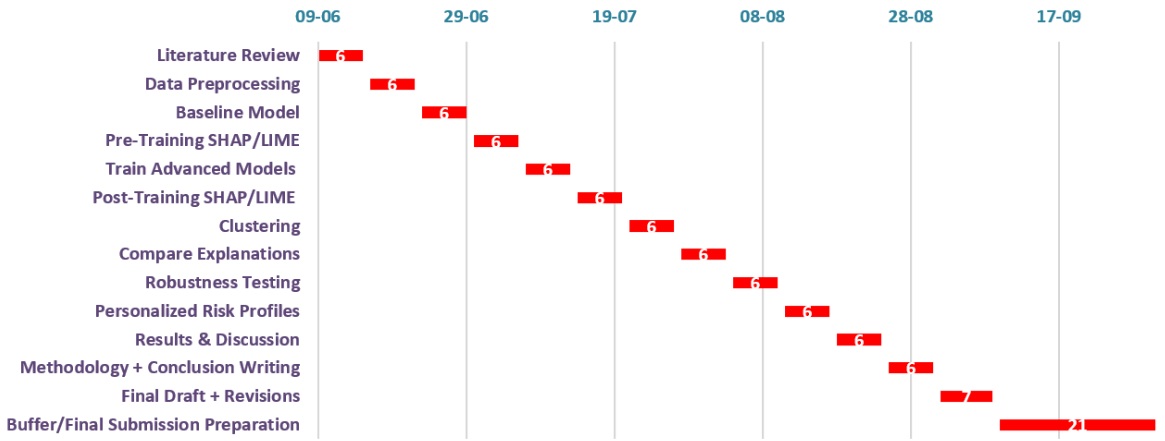

Figure 3:Gantt Chart describing the time line of my dissertation.

Risks

This project entails several potential risks that may affect its implementation and outcomes. As a data science student, I have assessed the following key project-specific risks, along with their estimated likelihood, potential impact, and proposed mitigation strategies. A risk register is provided below:

Risk Description	Likelihood	Impact	Mitigation Strategy

Data Quality Issues (e.g., missing or noisy data)	Medium	High	Apply robust data preprocessing; use imputation and feature selection techniques
Model Overfitting or Poor Generalisation	Medium	High	Use cross-validation; apply regularization; monitor evaluation metrics closely
Interpretability Confusion (complex SHAP/LIME outputs)	Medium	Medium	Provide visual explanations and narrative summaries; use simpler models when needed
Ethical Misuse or Misinterpretation of AI Outputs	Low	High	Include clear disclaimers; interpret results in an educational context only
Clustering Not Producing Useful Personas	Medium	Medium	Adjust number of clusters; evaluate using silhouette scores and domain knowledge
Instability in Explanations (adversarial attacks)	High	Medium	Measure robustness of SHAP/LIME with controlled feature perturbations
Technical Errors in Implementation or Tool Compatibility	Low	Medium	Use tested libraries (scikit-learn, SHAP, LIME); develop iteratively and test early
Time Constraints and Submission Deadlines	Medium	High	Follow the Gantt chart plan strictly; allocate buffer time for final preparation

Table 1:Describing the Risk

All risks are considered in light of ethical standards outlined in the Ethics Review Form. As no personally identifiable information will be used, the study presents minimal ethical risk and complies with institutional data protection policies. Potential user impact is mitigated by ensuring results are used exclusively for academic insight and improvement.

References

Choi, E., Joo, Y., & Kim, H. (2023) *Interpretable dropout prediction: Towards XAI-based personalized intervention*, Journal of Educational Data Mining, [online] Available at: https://link.springer.com/article/10.1007/s40593-023-00331-8 [Accessed 15 May 2025].

Gunasekara, D., & Saarela, M. (2025) *An explainable machine learning approach for student dropout prediction*, Expert Systems with Applications, [online] Available at: https://www.sciencedirect.com/science/article/abs/pii/S0957417423014355 [Accessed 15 May 2025].

Lopez, C., Ruan, S., Zhan, Y. and Wu, J. (2024) *A novel AI-driven model for student dropout risk analysis with explainable AI insights*, Discover Artificial Intelligence, [online] Available at: https://www.sciencedirect.com/science/article/pii/S2666920X24001553 [Accessed 15 May 2025].

Nguyen, H.T., Yu, J. and Pan, R. (2024) *Interpret3C: Interpretable student clustering through individualized feature selection*, arXiv preprint. Available at: https://arxiv.org/abs/2407.11979 [Accessed 15 May 2025].

Ribeiro, F., Lopes, F., & Silva, P. (2022) *On the use of explainable artificial intelligence to evaluate school dropout*, Education Sciences, 12(12), pp.1–16. Available at: https://www.mdpi.com/2227-7102/12/12/845 [Accessed 15 May 2025].

Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., Radi, N. and Kyriacou, T. (2019) *Machine learning approaches to predict student academic performance using educational data mining*, Information, 10(9), p.292. Available at: https://www.mdpi.com/2078-2489/10/9/292 [Accessed 15 May 2025].

Kaur, J. and Singh, M. (2022) *A hybrid approach for dropout prediction using machine learning and XAI*, Procedia Computer Science, 198, pp.151–158. Available at: https://www.sciencedirect.com/science/article/pii/S1877050922010262 [Accessed 15 May 2025].

Ullah, A., Asghar, S., & Al-Obaidy, S. (2021) *Clustering and classification techniques in educational data mining: A survey*, Education and Information Technologies, 26, pp.1507–1546. Available at: https://link.springer.com/article/10.1007/s10639-020-10337-1 [Accessed 15 May 2025].

Lundberg, S.M. and Lee, S.-I. (2017) *A unified approach to interpreting model predictions*, Advances in Neural Information Processing Systems (NeurIPS), 30. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf [Accessed 15 May 2025].

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) *“Why should I trust you?” Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp.1135–1144. Available at: https://dl.acm.org/doi/10.1145/2939672.2939778 [Accessed 15 May 2025].

Research Ethics Review Form for MSc Projects

Computer Science Research Ethics Committee (CSREC)

https://www.city.ac.uk/about/governance/committees/cs-research-ethics

Postgraduate students undertaking their final project in the Department of Computer Science must consider the ethics of their project work and ensure that it complies with research ethics guidelines and the law for data protection. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

To ensure that they give appropriate consideration to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

PART A: Ethics Checklist. All students must complete this part.The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

Part A: Ethics Checklist

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://researchmanager.city.ac.uk/. This type of research is not covered by City's process, and external approval from an appropriate institution is required.		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? e.g. because you are recruiting current NHS patients or staff? If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/	**NO**
1.2	Will you recruit participants who are covered by the Mental Capacity Act 2005? Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/	**NO**
1.3	Will you recruit any participants who are covered by the Criminal Justice System, for example, people on remand, prisoners and those on probation? Such research needs to be authorised by the ethics approval system of the National Offender Management Service.	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - **https://researchmanager.city.ac.uk/**		*Delete as appropriate*
2.1	Does your research involve participants who are unable to give informed consent? For example, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about protected characteristics (as identified by the Equality Act 2010)? *For example, to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**

2.6	Does your research involve invasive or intrusive procedures? These may include, but are not limited to, electrical stimulation, heat, cold or bruising.	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the Senate Research Ethics Committee (SREC), you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://researchmanager.city.ac.uk/. Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? For example, students studying on a particular course or module. If yes, then approval is also required from the Head of Department or Programme Director.	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**

A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**