Wassim Ben Youssef
Student ID: 160043583
Research Project Proposal
Supervisor: Dr. Artur D'Avila Garcez

Rule extraction from Deep Belief Networks applied to Visual Genome

Abstract- Famous for its great performances for pattern recognition, deep learning plays a key role in many domains and research areas such as medicine, finance, defence, robotics and aeronautics. The complexity of deep learning algorithms makes their efficiency but also their limit. Reasoning and processes of these algorithms are often incomprehensible despite their high-quality results. In this proposal, we expose a project that will lead to a system providing a better understanding and an improvement of restricted Boltzmann machines and deep belief networks.

Introduction

The aim of this project is to implement a system that helps to understand the running of a deep belief network which was conceived in [1]. This system is based on knowledge extraction and this knowledge is represented by logical rules. The system also provides a method to use the knowledge extracted to improve Deep Belief Networks. This implementation will then have to be tested on a specific dataset and the results will be analysed.

So, the products that will be generated are a knowledge extraction and insertion library and an analysis of its application on the Visual Genome dataset. The several stages of the research will be:

- Implementation of rule extraction for Restricted Boltzmann Machines
- Implementation of rule insertion for Restricted Boltzmann Machines
- Implementation of rule extraction for Deep Belief Networks
- Implementation of rule insertion for Deep Belief Networks
- Application of rule extraction and insertion for Deep Belief Networks to Visual Genome dataset

The research question for this project is:

How can I extract knowledge from Deep Belief Networks using logical rules, how can I use the knowledge extracted for predictions and how can I apply it to the Visual Genome Dataset?

The principal beneficiaries of this work:

- The Research Centre for Machine Learning of City, University of London for which the members published most of the papers dealing with knowledge extraction for deep learning and neural networks
- Artificial Intelligence industry which needs a better understanding of the behavior of its systems, especially in fields like Defense, Medicine, Law…
- Scientists and researchers working in explanation of Artificial Intelligence, especially for Deep Learning and Neural Networks (Explainable AI, The Defense Advanced Research Projects Agency …)

Critical Context

The need of Knowledge extraction

Artificial Intelligence allows to create systems that learn and act by following and imitating the behaviour of humans. Artificial Intelligence and especially Machine Learning and Deep Learning have spread out their wings and are nowadays applied in most of our daily tools while having a predominant role in many fields like Military Defence, Health Research or Finance. However, the processes followed by these machines often call out complicated mathematical models while dealing with a huge amount of information. The power and the effectiveness of these systems are then limited by their ability to make understand their reasoning and their logic to humans [3].

Therefore, Deep Learning systems and especially Artificial Neural Networks are famous for their high level of learning performance particularly in pattern and voice recognition. But the learning process is often assimilated to a black-box ([4] and [5]) unable to provide justifications for the results which can be crucially required in some areas of application. Hence, if we take the case of a neural network trained to classify pictures of animals, when the network recognizes an image as a cat, we have absolutely no idea why the system has converged to this conclusion, it could be because of the shape of the eyes, the ears or something else. The same issue can be raised with an image of tumour and a neural network which identify if the tumour is benign or malign. In this case, the doctor or the researcher would need to know the reasoning of the system. So this can be very confusing for the user who need to understand the system's decision in order to know when it fails and when it is reliable.

However, this lack of interpretability was not an issue with the first AI systems which used to reason with logical inference, providing a feedback of their inference process which can then be understood by humans [3]. These systems were much less efficient than the current ones though. But most of the new works and papers dealing with the subject of knowledge extraction on the new techniques of Artificial Neural Network are inspired by the logical process used in these earlier systems [5].

Algorithms and methods for knowledge extractions

Decision trees

Various first attempts to extract rules from neural network require specific assumptions and specific structures for the network [6]. That is why Schmitz et al [6] has implemented an algorithm that extracts binary decision trees from a trained neural network without any assumptions made on the network. This approach is specific to feedforward neural networks dealing with supervised learning. A simple definition of the decision tree would be "it is an analysis diagram, which can help decision makers to decide which is the best option between different options, by projecting possible outcomes. The decision tree, gives the decision maker an overview of the multiple stages by that will follow each possible decision" [7]. Let's take the example of an individual who wants to choose between going to a party or no:

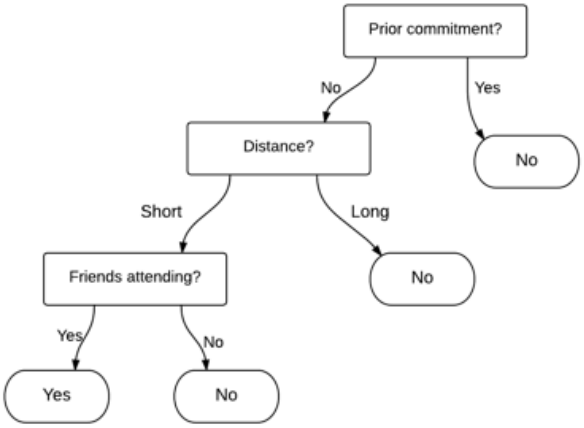

Figure 1: Binary decision tree, each node defines a test on values of property, terminal nodes are Boolean value of the goal class

The TREPAN algorithm was the first one developed to create decision trees that mimics the behavior of ANN ([4] and [5]).

Symbolic knowledge extraction

Another way to extract knowledge from the Black-Box of the network is to provide logic rules used as hypothesis ([1], [4], [13], [14]). In the hidden part (hidden layers) of the network, each hidden node represents a hypothesis about a specific rule. The hidden node will calculate the probability that "the rule implies a certain relation in the beliefs *b* being observed in the visible layer *V*, given the previously applied rules" [8]. For instance, a logic rule can be "The light is on and the door is not open", which is represented by the formula: $h \leftrightarrow x_1 \wedge \neg x_2$. That means that the hypothesis is true when x_1 (the light is on) is true and when x_2 (the door is open) is false. These kinds of rules are easily understandable by the user and they are based on the observations and data provided as inputs.

Rule extraction for Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are popular models mostly used for a data with a dynamic and temporal behavior like Time Series. Their particularity is that they have at least one feedback loop, which means that one of the output of one of the layers becomes an input of a previous layer [10]. To illustrate this concept, we take the example of a network with two layers each one with two neurons. One of the neuron of the second layer is connected to one of the neuron of the first layer via a feedback loop:

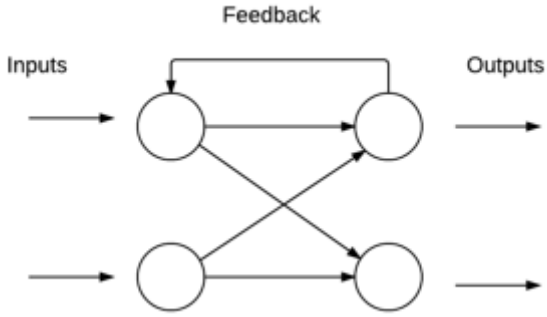

Figure 2: Illustration of recurrent neural network

A vast variety of RNN structures have been developed but only a few of them have been subject of works on symbolic knowledge extraction [9]. Most of them are listed by H. Jacobsson [9], and they use several kinds of logic rules (propositional logic, nonconventional logic, First-order logic and Finite state machines).

Unsupervised learning with Deep Belief Network and Restricted Boltzmann Machines

Unsupervised learning

The algorithms saw previously use supervised learning, instead of Restricted Boltzmann Machines (RBM) and Deep Belief Networks (DBN) that are unsupervised algorithms. Supervised learning means that we try to find information from data set with couples of "inputs-outputs". We then know what is the structure of the output, it can be a binary variable representing the answer of a question like "is this individual sick or no?". On the contrary, with supervised learning, we do not really know what we are seeking. The algorithm will then try to find by itself similarities in the data.

Restricted Boltzmann Machines and Deep Belief Network

The Restricted Boltzmann Machine is an unsupervised learning algorithm that learns to probabilistically rebuild the input, so that it tries to find a probability distribution of the data. The RBM is composed of one visible layer and one hidden layer without connections of neurons of the same layer and all visible nodes pass through all hidden nodes. The algorithm processes in two steps. First, there is a propagation from visible neurons to hidden neurons which apply an activate function on the values received from the visible neurons. Then the activations are the new inputs for the hidden layers which propagate in a backward pass through the visible layer:

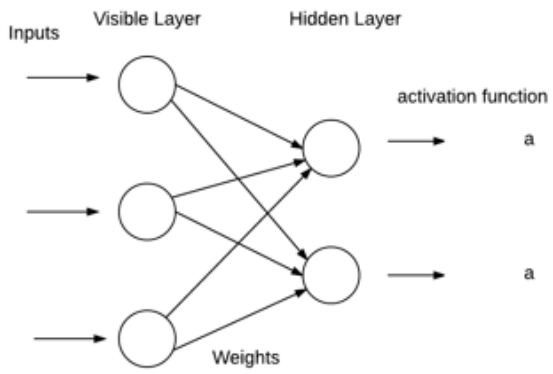

Figure 3: First step of RBM, propagation

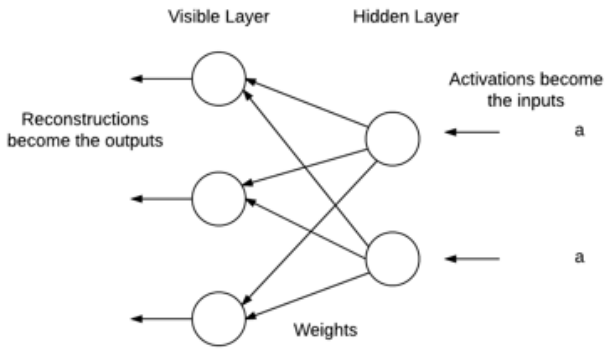

Figure 4: Second step, reconstruction

The Deep Belief Networks are a stack of RBMs learned one at a time [15]. The activations of the first hidden layer are hence used as inputs for the second layer, and so on until the last hidden layer from which the reconstruction step will began. A propagation will then occur until the first layer:

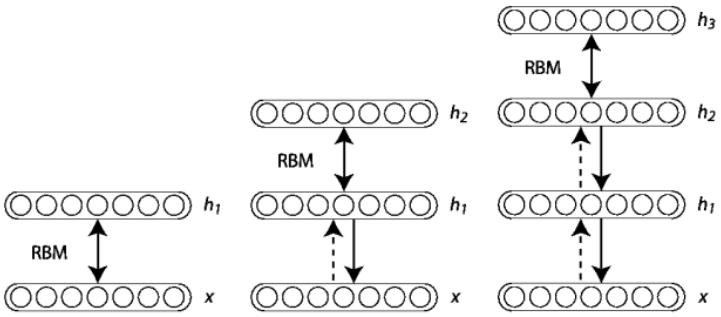

Figure 5: Illustration of Deep Belief Networks [12]

Rule extraction method for RBM and DBN and knowledge insertion

Logic rules extraction from deep networks as RBM and DBN is a field that have not been explored much and the first algorithm was designed in [14]. The only method investigated is based on confidence values and it is an extension of the Penalty Logic designed by G. Pinkas [16]. Previously, we saw that symbolic knowledge extraction is the concept of extraction of logic rules for each hidden node of a network. An algorithm was then conceived to extract the logic rules from an RBM, then it was extended for DBN. The rules are built by capturing the associations between input variables having strong connections through each hidden unit [1]. However, for a deep network with hundreds of nodes, that can makes a too high number of rules. So an inference algorithm was conceived to calculate confidence values for each rule extracted in order to hierarchize them. An experiment also show that the confidence values used as inputs for a classification problem can provide good results. Moreover, these confidence rules are also used to modify the weights between nodes of the network to improve it. This is what we can call knowledge insertion.

Approaches: Methods & Tools for Analysis & Evaluation

Literature survey

The research question was reflected and decided with the supervisor. A first literature review was done with the City University library, the Research Gate website and Google Scholar to have an overall view of the topic. Another literature search and review might be performed at the beginning of the project to have a better understanding of some aspects.

Analyze and taking in hand previous work

We will need to understand well the several algorithms conceived in [1]. We will also need to analyze the code of this work (provided by the supervisor) to understand how it works. This code is implemented in Matlab. A work must also be done on the several datasets used for this work to understand how they were handled and used for knowledge extraction and we need to understand the results. These datasets are: TiCC, MNIST, Yale and the DNA promoter problem. Including the literature survey, that part should take one week.

Implementation for RBM

Firstly, we transpose the Matlab code on rule extraction for RBM in Python. We will then run it on the DNA problem dataset, as done in [1], and evaluate the results, all of that in a period of approximately two weeks.

Two other weeks will be allocated to transpose the Matlab code on knowledge insertion for RBMs, to run it on TiCC, MNIST and Yale datasets and to evaluate and analyze the results.

Implementation for DBN

The same process will be followed for DBN. This part should be quicker given that it is an extension of the RBM procedure. Less than two weeks are allocated to knowledge extraction which will be tested and evaluated on MNIST dataset and the same time will be used for knowledge insertion.

Application to Visual Genome Dataset

We will use the Visual Genome dataset [17] to apply the extraction and insertion of knowledge for DBN implemented in Python. The Visual Genome dataset provides description of images by mapping the images to humans, human actions and objects with scene graphs. The dataset contains dense annotations of objects, attributes, and relationships between objects. The dataset also contains unconstrained image descriptions and question-answers about what is happening in the image. The Visual Genome dataset consists of seven main components: region descriptions, objects, attributes, relationships, region graphs, scene graphs, and question-answer pairs [18].

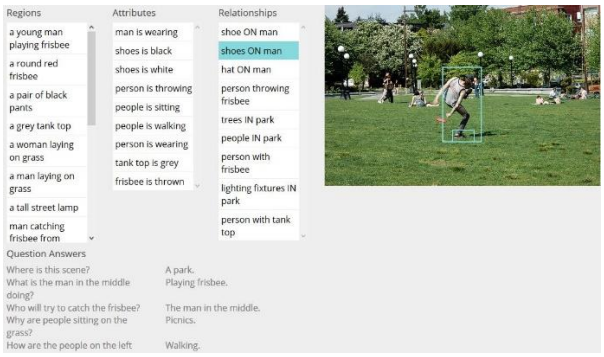

Figure 6: Representation of all the components detected in an image

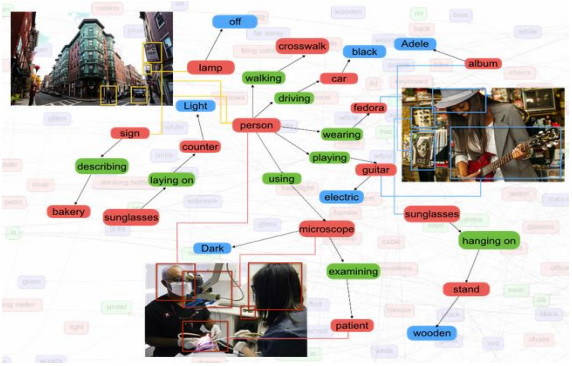

Figure 8: example of scene graphs for 3 images

The dataset contains 108,077 images. A work must be done to identify specifically what work will be done on the dataset and to understand how to work and handle with this dataset. A specific classification problem must be identified. We might follow the tutorial given in the website and analyze the documentation provided.

We will apply DBMs extraction and insertion knowledge library implemented before on this dataset and we will analyze the results. All this work will take less than two weeks.

Evaluation

To evaluate the models built for classification tasks with knowledge insertion, we will use a test average accuracy. We will compare the results of a classification problem for the networks without knowledge insertion and with knowledge insertion. Each dataset will be divided into training, validation and test sets, then each model will be run 10 times for each set and we will compute the average accuracy obtained. The two first sets will be used to set the parameters and the test set will be used for final evaluation. For the datasets used in [1], we will compare the results with the ones obtained in the paper. For Visual Genome dataset, the same process will be followed. So, a first phase of evaluation will be done for RBM, a second one for DBN and a last one for DBN applied to Visual Genome. All this job will be done in Python.

Complete report

The report will be started at the beginning of the project and will be completed at each step. We will begin for when the first implementation and evaluation of RBMs algorithms and applications will be done. One week will be allocated to that. We will then continue to write for one week after completing implementation for DBNs. We will focus only on the report for the last three weeks of the project.

Project Meetings

Several short meetings would be scheduled if issues are encountered and if needed. Three formal meetings will be scheduled at the end of each key step of the project to present formally what have been done.

Work Plan

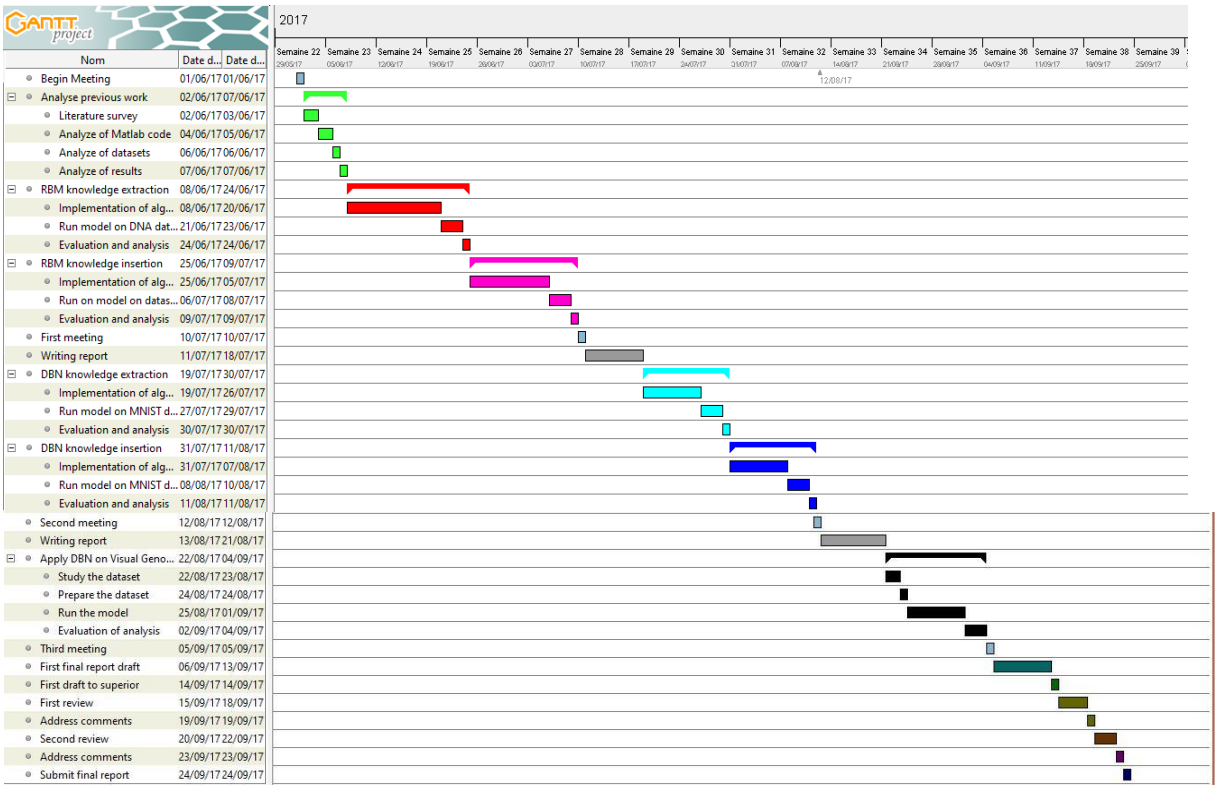

Figure 9: project timeline

Risks

Description	Likelihood (1-5)	Consequence (1-5)	Impact (L x C)	Mitigation
A library used in Matlab code is not available in Python	2	5	10	Implement it by ourselves if it is easy, else we must find another solution
Existing results cannot be replicated exactly	4	2	8	Make a relative comparison
Supervisor is busy and cannot do any additional meetings others than the ones scheduled at the beginning	4	3	12	Take more time for the formal meetings and communicate by e-mails
Some networks take too time to run	3	4	12	Reduce the number of hidden units or/and hidden layers
Code takes a lot of time to run in my laptop	3	4	12	Use City University's clusters
Code is accidentally lost	1	5	5	Use GitHub to keep an updated version online

References

[1] S. N. Tran and A. D'Avila Garcez, "Deep Logic Network: Inserting and Extracting Knowledge from Deep Belief Networks," in *IEEE Transactions on Neural Networks and Learning Systems*, PP(99), 08 November 2016

[2] W. Knight, "The U.S. Military Wants Its Autonomous Machines to Explain Themselves" in *MIT Technology Review, Intelligent Machines*, March 2017. Available: https://www.technologyreview.com/s/603795/the-us-military-wants-its-autonomous-machines-to-explain-themselves/ . [Accessed 14 April 2017]

[3] D. Gunner, "Explainable Artificial Intelligence (XAI)" in *Broad Agency Announcement*, DARPA-BAA-16-53, August 10, 2016

[4] C. Percy, A. D'Avila Garcez, S. Dragicevic, M. V. M. França, G.G. Slabaugh and T. Weyde, "The Need for Knowledge Extraction: Understanding Harmful Gambling Behavior with Neural Networks," in *Frontiers in Artificial Intelligence and Applications*, 285, pp. 974-981. doi: 10.3233/978-1-61499-672-9-974, 2016

[5] M. Rangwala and G.R. Weckman, "Extracting Rules from Artificial Neural Networks utilizing TREPAN" in *IIE Annual Conference.Proceedings*, pp. 1-6, 2006

[6] G. P. J. Schmitz, C. Aldrich, and F. S. Gouws, "ANN-DT: An Algorithm for Extraction of Decision Trees from Artificial Neural Networks", in *IEEE Transactions on neural networks, Vol. 10, no. 6, November 1999*

[7] N. I. Khawaldah and N. A. Ishtayeh, "Binary Decision Tree", *Al Zaraqa University*

[8] H.L.H. (Leo) de Penning, A. d'Avila Garcez, L. C. Lamb3 and J-J C. Meyer, "A Neural-Symbolic Cognitive Agent for Online Learning and Reasoning,", in *Conference: IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Barcelona, Catalonia, Spain, July 16-22, 2011

[9] H. Jacobsson, "Rule Extraction from Recurrent Neural Networks: A Taxonomy and Review," in *Neural Computation* 17, 1223–1263**,** 2005

[10] S. Haykin, "Neural Networks and Learning Machines", 3rd edition, Prentice Hall, Upper Saddle River, N.J., 2009

[11] Skymind, "https://deeplearning4j.org/restrictedboltzmannmachine", 2017. [Accessed 14 April 2017]

[12] Y. Bengio, "Learning Deep Architectures for AI," in *Preliminary version of journal article with the same title appearing in Foundations and Trends in Machine Learning*, 2009

[13] S. N. Tran and A. d'Avila Garcez, "Logic Extraction from Deep Belief Network", in *ICML 2012 Representation LearningWorkshop*, Edinburgh, July 2012.

[14] S. N. Tran and A. d'Avila Garcez, "Knowledge Extraction from Deep Belief Networks for Images"

[15] P. Smolensky. "Information processing in dynamical systems: Foundations of harmony theory," in Rumelhart, D. E. and McClelland, J. L., editors, Parallel Distributed Processing: Volume 1: Foundations, pages 194–281. MIT Press, Cambridge, 1986.

[16] G. Pinkas. Reasoning, "Nonmonotonicity and learning in connectionist networks that capture propositional knowledge" in Artificial Intelligence, 77(2):203 –247, 1995.

[17] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332, 2016

[18] R. Krishna, *Master Thesis: Visual Genome Crowdsourced Visual Knowledge Representation*, Stanford University, March 2016

Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. 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.

In order to ensure that appropriate consideration is given 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 B: Ethics Proportionate Review Form. Students who have answered "no" to questions 1 – 18 and "yes" to question 19 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in this case. The approval may be provisional: the student may need to seek additional approval from the supervisor as the project progresses.

A.1 If your answer to any of the following questions (1 – 3) is YES, you must apply to an appropriate external ethics committee for approval.		*Delete as appropriate*
1.	Does your project require approval from the National Research Ethics Service (NRES)? For example, because you are recruiting current NHS patients or staff? If you are unsure, please check at http://www.hra.nhs.uk/research-community/before-you-apply/determine-which-review-body-approvals-are-required/.	**No**
2.	Does your project involve participants who are covered by the Mental Capacity Act? If so, you will need approval from an external ethics committee such as NRES or the Social Care Research Ethics Committee http://www.scie.org.uk/research/ethics-committee/.	**No**
3.	Does your project involve participants who are currently under the auspices of the Criminal Justice System? For example, but not limited to, people on remand, prisoners and those on probation? If so, you will need approval from the ethics approval system of the National Offender Management Service.	**No**

A.2 If your answer to any of the following questions (4 – 11) is YES, you must apply to the City University Senate Research Ethics Committee (SREC) for approval (unless you are applying to an external ethics committee).		*Delete as appropriate*
4.	Does your project involve participants who are unable to give informed consent? For example, but not limited to, 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**
5.	Is there a risk that your project might lead to disclosures from participants concerning their involvement in illegal activities?	**No**
6.	Is there a risk that obscene and or illegal material may need to be accessed for your project (including online content and other material)?	**No**

7.	Does your project involve participants disclosing information about sensitive subjects? For example, but not limited to, health status, sexual behaviour, political behaviour, domestic violence.	**No**
8.	Does your project involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning? (See http://www.fco.gov.uk/en/)	**No**
9.	Does your project involve physically invasive or intrusive procedures? For example, these may include, but are not limited to, electrical stimulation, heat, cold or bruising.	**No**
10.	Does your project involve animals?	**No**
11.	Does your project involve the administration of drugs, placebos or other substances to study participants?	**No**

A.3 If your answer to any of the following questions (12 – 18) is YES, you must submit a full application to the Computer Science Research Ethics Committee (CSREC) for approval (unless you are applying to an external ethics committee or the Senate Research Ethics Committee). Your application may be referred to the Senate Research Ethics Committee.		*Delete as appropriate*
12.	Does your project involve participants who are under the age of 18?	**No**
13.	Does your project 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**
14.	Does your project involve participants who are recruited because they are staff or students of City University London? For example, students studying on a specific course or module. (If yes, approval is also required from the Head of Department or Programme Director.)	**No**
15.	Does your project involve intentional deception of participants?	**No**
16.	Does your project involve participants taking part without their informed consent?	**No**
17.	Does your project pose a risk to participants or other individuals greater than that in normal working life?	**No**
18.	Does your project pose a risk to you, the researcher, greater than that in normal working life?	**No**

A.4 If your answer to the following question (19) is YES and your answer to all questions 1 – 18 is NO, you must complete part B of this form.		
19.	Does your project involve human participants or their identifiable personal data? For example, as interviewees, respondents to a survey or participants in testing.	**No**