March 2018

Research Methods and Professional Issues

INM373

Research Proposal - Task II

Incomplete Information Problem using Reinforcement Learning and Opponent Modelling

MSc. Data Science – PT II

Academic Year 2017-2018

Prof. Ernesto Priego

CALVO Mathieu

CITY UNIVERSITY LONDON

Table of Contents

Table of Contents

1. **Purpose**

Introduction

This research project will aim at tackling an incomplete information game using reinforcement learning techniques.

Broadly speaking, Reinforcement learning (RL), in the context of Artificial Intelligence (A.I.), is an approach of Machine Learning (ML) that trains algorithms using a system of reward and punishment. An agent trained using RL learns by interacting with its environment, and progressively grasp what actions yield positive outcomes on the long run. It differs from supervised ML techniques in that it does not teach explicitly the agent how to perform a task, but rather create the conditions for it to work through the problem on its own.

Studying games in the A.I. field is an historical precedent. They have a framed environment governed by rules that constitute a good simplification of real-world problems. "For more than a half century, games have continued to act as test beds for new ideas and the resulting successes have marked important milestones in the progress of AI."[1].

A game is said to be a complete information game when all players have access to the same amount of information and the full state of the environment is visible to all. In other words, each player knows or can see what the other player is doing. The most commonly used example of a complete information game is chess, where all the available information needed to take the most appropriate decision is available and visible on the board at every single moment.

By contrast, an incomplete information game is a game where the environment is only partially observable to all players, "i.e. games where at least one player is uncertain about another player's payoff function" [4]. In many card games for example, players have access to privately-owned and hidden cards, they typically hold their private cards (also known as "hands") face-down to make sure nobody access the information they contain. In that context, each player cannot see directly, nor infer with certainty, what cards the opponents hold but nonetheless needs to take actions in this environment characterized by its inherent uncertainty. "Such games are strategically challenging. A player has to reason about what others' actions signal about their knowledge. Conversely, the player has to be careful about not signalling too much about her own knowledge to others through her actions". [3]

This privately owned information has been and continues to be a great challenge for the elaboration of efficient RL algorithms and, historically, the trained A.I. agents have been consistently dominated by their human counterparts and therefore constituted a new A.I. frontier for a long time. "Poker has emerged as a standard benchmark in this space" [3].

In the literature, the terms “imperfect information” and “incomplete information” are sometimes used interchangeably to qualify a game. And the game of poker, especially, is described as a game of “imperfect knowledge” [6] as much as an “incomplete information game” [3]. The frontier between the two concepts seem quite thin, evolving and most of the time ignored by the practitioners so we will consider the terms as interchangeable in this report.

The poker game

In this project, we initially aim at targeting arguably the most popular incomplete information card game at the moment, the very famous game of poker. Specifically, we will only consider its most common variant, also known as the Limit Texas Hold’em Poker game, which is played with two private cards face down dealt per player and five community cards dealt in the middle of the table, face up.

The best combination of five cards into the seven wins (private plus community cards). There is a ranking of the combinations to help determine the winner. The community cards are dealt progressively, and their dealing is separated by betting rounds and conditional on the players implicitly agreeing on the amount to bet at every single round. The hand can terminate in two ways. The first option is that at least two players agree on the amount to bet until the end and everybody see private cards owned by those players present at the end of the hand (the so-called “showdown”) and we determine who has the best combination by referring to the ranking. The second option is that one player is willing to bet more than the others and therefore wins before the final stage, the other players who abandon the hand are said to be “folding”.

“The challenges introduced by poker are many. The game involves a number of forms of uncertainty, including stochastic dynamics from a shuffled deck, imperfect information due to the opponent’s private cards, and, finally, an unknown opponent. These uncertainties are individually difficult and together the difficulties only escalate. A related challenge is the problem of folded hands, which amount to partial observations of the opponent’s decision making contexts. (…). A third key challenge is the high variance of payoffs, also known as luck. This makes it difficult for a program to even assess its performance over short periods of time. To aggravate this difficulty, play against human opponents is necessarily limited. If no more than two or three hundred hands are to be played in total, opponent modelling must be effective using only very small amounts of data. Finally, Texas hold’em is a very large game. It has on the order of 10^{18} states, which makes even straightforward calculations, such as best response, non-trivial.”[5]

Project scope

In order to limit the complexity of this research project, we will proceed to a few simplifications.

- We will limit our analysis to one versus one poker games, also known as "heads-up" poker games.
- Our poker-playing agent will specialize in playing games where the initial stakes do not change over the course of the session, also known as "cash-games".
- We will assume that only a limited number of bet sizes are possible at each point in time
- We will assume that each session ends after a given number of hands.
- And most importantly, the initial opponent of our poker-playing agent will be limited to a simple reflex-agent that follows a fixed policy

In real-life poker games, other players (or agents, to use the RL terminology) are dynamically recalibrating their policies by constantly adjusting their decision-making process and reacting to the information they gather about players. This brings about a new layer of complexity that falls out of the scope of this project. We will assume at first that our opponent policy is static and we will design it in such a manner.

The environment for the RL problem I will be finding a solution for is best described as follows:

- **Partially observable**: our agent can't see the other player's hand or even the remaining cards in the deck, but can see his private cards plus the community cards.
- **Stochastic**: luck or randomness is involved, as the deck of cards is shuffled randomly.
- **Sequential**: each decision takes part of several narratives, on one hand the hand being played (four betting rounds), and on the other hand the session being played (stack, history of hands and observations, etc...).
- **Static**: other agent is not constantly adjusting its decision process mechanism.
- **Discrete**: agents can only bet multiple of a given number (the big blind) up until the amounts of chips in their stack at the moment of the decision
- **Known**: finite universe of possible hands and situations. Confined universe where rules are known in advance, hands are ranked and sequence is set.

Our poker-playing agent will need to have the following characteristics in order to be able to solve the problem at hand:

- **Model-based agent**: it needs to know the probabilities inherent to the game, and update his perception of the opponent as observations arise
- **Goal-based agent**: it needs to maximize profit
- **Utility-based agent:** it also needs to maximize quality of its decisions, as correct decisions that lead to "bad beats" or unfortunate events, should also be rewarded. Poker is a high-luck high-skills game and agent needs to distinguish process from short-term results to a great extent.

The emphasis of this project will be on training an agent on an incomplete-information game using both RL and opponent modelling techniques. Ultimately, we want our agent to learn the weaknesses in his opponent fixed policy, what his actions says about the private information he has and, consequently, to respond by taking the correct actions to harness those weaknesses efficiently.

Project outcome

The project results will be of interest to the A.I. community with Imperfect-Information Games being at the very centre of attention in the recent years.

Notably, the AAAI Workshop on AI for Imperfect-Information Games is a "forum where researchers studying theoretical and practical aspects of imperfect-information games can share current research and gather ideas about how to improve the state of the art and advance AI research in this area." [12]. The association organises an "Annual Computer Poker Competition" that also welcomes reports on algorithms for solving large imperfect information games, regardless of the aspects being tackled, ranging from game theory to opponent modelling. "Every year researchers and hobbyists interested in computational poker gather at this workshop to discuss in person the latest research and results." [12].

Some companies, like Leanpoker.org, organise coding events around poker-playing algorithms, where the goal is to improve the efficiency and performances of the agent with very little time to do so. "A Lean Poker event lasts for about 8 to 10 hours. During this time, self-organized teams develop and improve poker-playing robots." [13].

Any improvement in that field can undeniably have far-reaching consequences and be of interest to many, as "although seemingly playful, game theory has always been envisioned to have serious implications [e.g., its early impact on Cold War politics]. More recently, there has been a surge in game-theoretic applications involving security, including systems being deployed for airport checkpoints, air marshal scheduling, and coast guard patrolling" [1].

2. Critical context

The new milestone in A.I. research

"The best poker player in the world is a man-made robot named Libratus, as of 2017" [10], Poker Sites recently headlined in its "rise of machines against humans" article.

Its inventor, Dr. Tuomas Sandholm from the Carnegie Mellon University, wrote "Libratus, an AI that, in a 120,000-hand competition, defeated four top human specialist professionals in heads-up no-limit Texas hold'em, the leading benchmark and long-standing challenge problem in imperfect-information game solving." [9]. Indeed, Poker has become the leading benchmark for solving incomplete information problem and the recent news is indeed of a significant importance.

The A.I. community has gone through many breakthroughs in the recent years, with IBM's DeepBlue chess-playing agent and Google's AlphaGo Go-playing agent most notably, both winning against the very best humans in their respective field of expertise.

Libratus is perhaps of an even greater significance, because no nontrivial imperfect information game had ever been dominated by a non-human intelligence, also because of its extraordinary computational cost associated with its algorithm that was "powered by the Bridges system, a high-performance computer at the Pittsburgh Supercomputer Center" [11]. Sandholm believes it is "further pushing the boundary of what is computationally possible, and facilitates adoption of these approaches to additional games of importance, for example, negotiation, auctions, and various security applications." [3].

Opponent modelling at the very heart

In a conversation recounted by Bronowski, von Neumann, the founder of modern game theory, made the observation that: "Real life consists of bluffing, of little tactics of deception, of asking yourself what is the other man going to think I mean to do. And that is what games are about in my theory" [14].

Indeed, the very first designed poker-playing agents were lacking this capacity of adapting to humans' strategies and were quickly outsmarted. "Our initial experience with a poker-playing program was positive (Billings et al. 1997). However, we quickly discovered how adaptive human players were. In games played over the Internet, our program, Loki, would perform quite well initially. Some opponents would detect patterns and weaknesses in the program's play, and they would alter their strategy to exploit them. One cannot be a strong poker player without modelling your opponent's play and adjusting to it." [6].

In poker, as in many imperfect information games and real-life situations, being able to model one's opponent decision-making process is a decisive factor for success, and often makes the difference between winning and losing. "In strategic games like chess, the performance loss by ignoring opponent modelling is small, and hence it is usually ignored. In contrast, not only does

opponent modelling have tremendous value in poker, it can be the distinguishing feature between players at different skill levels. If a set of players all have a comparable knowledge of poker fundamentals, the ability to alter decisions based on an accurate model of the opponent may have a greater impact on success than any other strategic principle." [6].

As Lockett and Miikkulainen brilliantly put it "the opponent's behaviour provides the primary window into the opponent's state" [7] and assessing the strength of your opponent's hand will necessarily require analysing in depth your opponent's behaviour, not only in the hand being played but also during the session or even since first encounter. One should also be mindful of how the opponent will model one's actions, and how he has done so in the past.

Several models have been tried, Billings used a statistical approach "to estimate the strength of the opponent's hand given his history of calling, raising, or folding" [7], some tried to use a common classification or categorization of the opponent strategy as tight, loose, passive or aggressive. But what seems to have been Sandholm's ground-breaking innovation is to exploit suboptimal opponents. "While playing an equilibrium guarantees at least the value of the game in a two-player zero-sum game, often much higher payoffs can be obtained by deviating from equilibrium to exploit opponents who make significant mistakes. For example, against a poker opponent who always folds, the strategy of always raising will perform far better than any equilibrium strategy (which will sometimes fold with bad hands)." [8]. His approach "combine game theoretic reasoning and pure opponent modelling, yielding a hybrid that can effectively exploit opponents after a small number of interactions" [8].

That's precisely those features that makes this new milestone in A.I. so important, and paves the way to many fascinating breakthroughs.

3. Approaches

The challenges

This research project will entail several challenges:

- Write a poker game framework programmatically, including flow control, card shuffling and the sequence of events.
- Retrieve the poker game statistical knowledge, with hand values plus hand potentials (forward-looking) at each betting round based on available information.
- Design a model of opponent's possible range of hands given its action during the hand and during the session. This will involve a matrix of all possible combinations (excluding impossible ones, inferred from visible cards). This should allow for a probability estimation of our opponent bluffing.
- Design a decision process model that take many factors into account: position, players' stacks (number of chips), hand value and hand potential, opponent modelling
- Maximize quality of decisions knowing that good decisions could often lead to losses, and especially because our agent will not necessarily know what was the quality of its decisions (because only the showdown reveals the full state of the environment)

I have found and bookmarked important open-code sources for many aspects of this problem and also relevant papers discussing those topics. There will be a challenge with the computational and memory cost associated with the amount of information our agent needs to be feeding from.

The plan

I plan to design the software framework entirely in python which is a high-level programming language I am proficient with. If computational power becomes an issue I plan to use a distributed cloud computing service like AWS to run my experiments.

No data is needed as our agent learns from the environment created and the repetition of epochs. The challenge will be in designing the building blocks and dealing with limited memory. Some abstractions and simplifications will need to be made in order to prune the spectrum of possibilities and make it manageable.

I plan to design a fixed-policy opponent that has access to the same statistical knowledge about the game as our agent and have him take decisions that have a positive expected value based on the assumption that he is against a random hand.

I will use Q-learning and Temporal Difference techniques to make my agent learn and build up knowledge during the training. My agent needs to be in a position to learn both about the game and the opponent. The exploration phase will inform about what random actions yield. The exploitation phase will need to leverage the lessons learnt during training. Ultimately, my agent

should be in a position to rely more on the weaknesses of the opponent than on its knowledge about the game and probabilities.

In the first discussion with my supervisor, I plan to discuss all these building blocks in detail and shore up the guidelines of the first phase of this project.

Performance metrics

The agent's performance will be evaluated after each session (or epoch). The session could end in two manners: either the pre-defined maximum number of hands has been played or one of the players loses the entire amount of its initial stack before then.

In both cases, the performance of the agent after the session is measured in terms of chips (the abstracted version of money in poker) won or lost.

We then observe the evolution of that metric while the agent learns through learning epochs. What we would like to observe is a significant increase in our winning rate as training unfolds, in such a way that our poker-playing becomes a positive expected-value player which is a player that will statistically win on the long-run.

Additionally, we will measure the capacity of our agent to depart from a pure environmental observation-based decision making to factor in a modelling of the opponent. In other words, we will evaluate how much our agent will rely on the assessment of his hand and its potential to take decisions versus relying on the assessment of the strength and behaviour of its opponent.

4. Work Plan

Several phases

In order to plan and manage the project properly, I have chosen to use a Gantt chart that will detail the sequence of tasks and milestones that I intend to go through, along with their respective expected time of delivery. This format is really useful as it let you see the overlap of processes and allows you to see the progress made and the dependencies between the processes.

The first phase of this project is shown in green and includes the elaboration of this research proposal, a comprehensive literature review, some initial planning and the very first checkpoint with my supervisor to discuss in depth and in details how I intend to tackle the subsequent tasks.

The second phase is shown in blue and will involve programming all the building blocks that I will need in order to complete my experiments, as mentioned in the previous sections: the poker game environment, the statistical modelling of hand value and potential, the opponent modelling and finally the decision making process. These items do not need to be handled sequentially and those tasks will overlap as I work on versioning them. Another checkpoint will be scheduled with my supervisor towards the end of this phase.

The third phase, in red in the graphical representation, will be dedicated to running experiments, varying parameters, analysing results and running new experiments. This phase will be first explorative in nature, and will lead to more appointments with my supervisor. These checkpoints will help me steer my experiments towards a more exploitative approach, in which I consolidate my results and orient them towards my project conclusions.

Finally, the purple phase will be dedicated to preparing the final report, with the final presentation and submission in mind.

Graphical representation

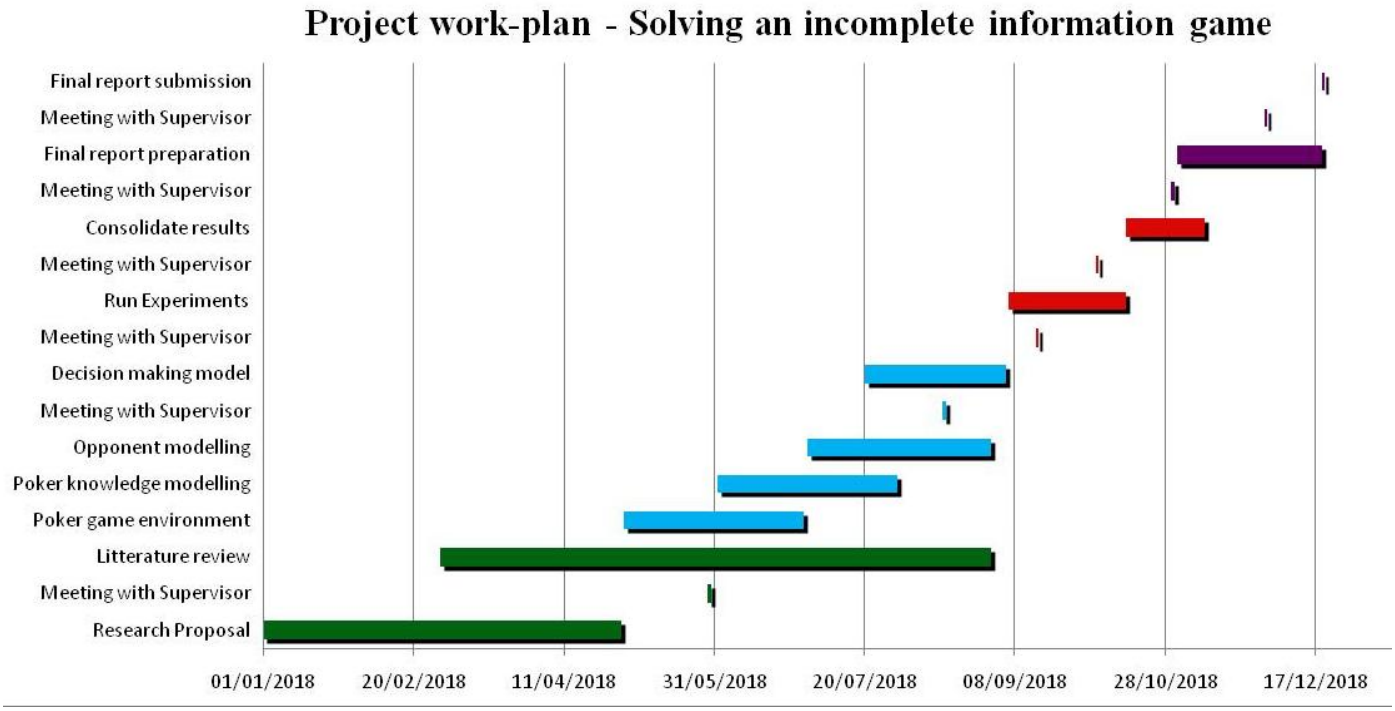

Figure 1. A graphical representation of the project work-plan using a Gantt chart

5. Risks

Rating for Likelihood and Seriousness for each risk			
L	Rated as Low	E	Rated as extremely serious
M	Rated as Medium	NA	Not Assessed
H	Rated as High		

Grade: Combined effect of Likelihood/Seriousness					
	Seriousness				
Likelihood		low	medium	high	EXTREME
	low	N	D	C	A
	medium	D	C	B	A
	high	C	B	A	A

Recommended actions for grades of risk	
Grade	**Risk mitigation actions**
A	Mitigation actions, to reduce the likelihood and seriousness, to be identified and implemented as soon as the project commences as a priority.
B	Mitigation actions, to reduce the likelihood and seriousness, to be identified and appropriate actions implemented during project execution.
C	Mitigation actions, to reduce the likelihood and seriousness, to be identified and costed for possible action if funds permit.
D	To be noted - no action is needed unless grading increases over time.
N	To be noted - no action is needed unless grading increases over time.

Description of Risk (including any identified 'triggers')	**Impact on Project** (Identify consequences [1])	Assessment of **Likelihood**	Assessment of **Seriousness**	**Grade**	**Mitigation Actions** (Preventative or Contingency)	**Timeline** for mitigation action(s)
Complexity of problem at hand becomes unmanageable with available resources because of a lack of time	Jeopardize ability to deliver in time	M	H	B	Preventive - Monitor closely evolution versus work-plan Contingency - Re-scope project: simplify problem at hand by choosing another game or simplifying existing	31/08/18
Computing complexity of problem at hand becomes unmanageable with available resources	Lack of computing power will stall progress and potentially affect delivery	H	H	A	Preventive - Estimate computing power needed at an early stage Contingency - Re-scope project: simplify problem at hand by choosing another game or simplifying existing	31/08/18
Inability to develop appropriate model: a building block turns out to be challenging from a point of view of computing memory	Lack of memory will affect ability to deliver	M	H	B	Contingency - Re-scope project: simplify problem at hand by choosing another game or simplifying existing	TBC
Running experiments turns out to be too challenging as the number of training epochs has to be substantial	Affect scope and results	H	M	D	Contingency - Use a cloud computing service like Amazon Web Services	TBC

6. References

1. Bowling, M., Burch, N., Johanson, M., & Tammelin, O. (2015). Heads-up limit hold'em poker is solved. Science, 347(6218), 145-149.
2. McCurley, P. (2009). An artificial intelligence agent for Texas hold'em poker. Undergraduate dissertation, University of Newcastle Upon Tyne.
3. Sandholm, T. (2010). The state of solving large incomplete-information games, and application to poker. AI Magazine, 31(4), 13-32.
4. Slantchev, B. L. (2008). Game Theory: Static and dynamic games of incomplete information. Dept of Political Science, Univ San Diego.
5. Southey, F., Bowling, M. P., Larson, B., Piccione, C., Burch, N., Billings, D., & Rayner, C. (2012). Bayes' bluff: Opponent modelling in poker. arXiv preprint arXiv:1207.1411.
6. Davidson, A., Billings, D., Schaeffer, J., & Szafron, D. (2000). Improved opponent modeling in poker. In International Conference on Artificial Intelligence, ICAI'00 (pp. 1467-1473).
7. Lockett, A. J., & Miikkulainen, R. (2008, December). Evolving opponent models for Texas hold'Em. In Computational Intelligence and Games, 2008. CIG'08. IEEE Symposium On(pp. 31-38). IEEE.
8. Ganzfried, S., & Sandholm, T. (2011, May). Game theory-based opponent modeling in large imperfect-information games. In The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2 (pp. 533-540). International Foundation for Autonomous Agents and Multiagent Systems.
9. Brown, N., & Sandholm, T. (2017). Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. Science, eaao1733.
10. Poker Sites. (2017, December 3). The rise of machines against humans. Retrieved from https://www.iafrikan.com/2017/12/03/the-frontier-of-ai-how-advanced-is-the-new-class-of-artificial-intelligence-in-2017/
11. Katyanna Quach (2017, December 19). The Revealed: How Libratus bot felted poker pros – and now it has cyber-security in its sights. Retrieved from https://www.theregister.co.uk/2017/12/19/poker_bot_libratus_ai/
12. Retrieved from http://www.cs.cmu.edu/~noamb/aaai18/workshop.html
13. Retrieved from https://leanpoker.org/how-it-works
14. J. Bronowski, The ascent of man, Documentary (1973). Episode 13

7. Ethics checklist

Part A. of Ethics review form for B.Sc, M.Sc and MA research projects

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 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, please check at http://www.hra.nhs.uk/research-community/before-you-apply/determine-which-review-body-approvals-are-required/)	**No**
2.	Will you recruit any participants who fall under the auspices of the Mental Capacity Act? (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**
3.	Will you recruit any 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? (Such research needs to be authorised by the ethics approval system of the National Offender Management Service.)	**No**

If your answer to any of the following questions (4 – 11) is YES, you must apply to the Senate Research Ethics Committee for approval (unless you are applying to an external ethics committee):		*Delete as appropriate*
4.	Does your research 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 research 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 research study (including online content and other material)?	**No**
7.	Does your research involve participants disclosing information about sensitive subjects?	**No**
8.	Does your research involve the researcher travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning? (http://www.fco.gov.uk/en/)	**No**
9.	Does your research involve invasive or intrusive procedures? For example, these may include but are not limited to, electrical stimulation, heat, cold or bruising.	**No**

10.	Does your research involve animals?	**No**
11.	Does your research involve the administration of drugs, placebos or other substances to study participants?	**No**

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 research involve participants who are under the age of 18?	**No**
13.	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**
14.	Does your research involve participants who are recruited because they are staff or students of City University London? For example, students studying on a particular course or module. (If yes, approval is also required from the Head of Department or Programme Director.)	**No**
15.	Does your research involve intentional deception of participants?	**No**
16.	Does your research involve participants taking part without their informed consent?	**No**
17.	Does your research pose a risk to participants greater than that in normal working life?	**No**
18.	Does your research pose a risk to you, the researcher(s), greater than that in normal working life?	**No**