UNIVERSITY OF LIVERPOOL

ELEC340 SPECIFICATION REPORT FOR PROJECT

Mapless Navigation with Deep Reinfocement Learning

Author:
Minghong Xu (201601082)

Supervisor:
Dr Murat Uney

Assessor:
Unassigned

Declaration of academic integrity

I confirm that I have read and understood the University's Academic Integrity Policy. I confirm that I have acted honestly, ethically and professionally in conduct leading to assessment for the programme of study. I confirm that I have not copied material from another source nor committed plagiarism nor fabricated, falsified or embellished data when completing the attached piece of work. I confirm that I have not copied material from another source, nor colluded with any other student in the preparation and production of this work.

Abstract

This report specify the project "Mapless Navigation with Deep Reinforcement Learning". The general plan is described through Gantt chart. Tasks are divided, with timeframes for completion and criteria for verifing the completion of tasks, and milestones and deliverables are defined. Motivation and methodology are ~~stated as well~~. The background to the project is presented in the Introduction, Literature Review, and Result section.

15th October 2022

Contents

List of Figures

List of Tables

1 Introduction

Reinforcement learning (RL) is a machine learning paradigm which leverages experiences gained through trial and error to optimise a function that implements a policy for a paticular agent in a paticular environment [1].

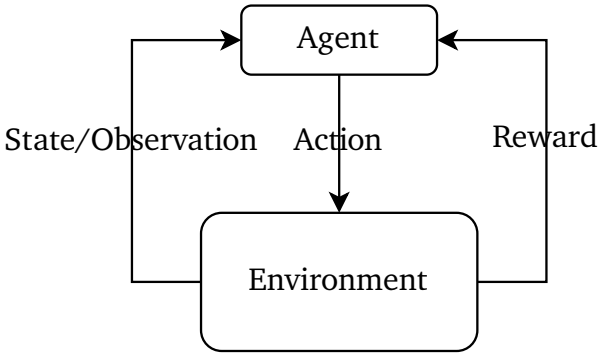


Figure 1: Agent-Environment interaction in RL

Deep reinforcement learning (DRL) is a type of RL which exploits deep neural network, an artificial neural network (ANN) composed of serveral layers, for function approximation [1].

Intuitively, agent learns how to act by iteratively taking an action and receiving a reward. Formally, this process is modelled as Markov decision process (MDP) which is a quintuple, $\langle S, A, R, P, \rho_0 \rangle$, where

- S, called state space, is the set of all valid states,
- A, called action space, is the set of all valid actions,
- R is the reward function whose input can be state, state-action pair, or state, action, and next state,
- P is the transition probability function represents the probaility of transitioning into next state if take a certain action in current state,
- and ρ_0 is the distribution of initial state [1].

The problem is how to find an optimal policy which maximises cumulative reward accumulates through each state.

$$\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t} R(s_t, a_t) \right]$$

Many algorithms have been proposed to try to solve this problem. These algorithms can be broadly divided into

- model-based which utilises environment model that predicts state transitions and rewards
- and model-free which based on common case that the model is not available to agent.

Two main approaches of model-free algorithm are Policy Optimisation and Q-Learning. Policy Optimisation is stable and reliable but sample ineffient, while Q-Learning is sample effient but is not principled so more likely to fail. Of the three state-of-the-art model-free DRL algorithms, PPO, TD3, and SAC, PPO uses only the first approach, while the latter two are a mixture of both apporaches.

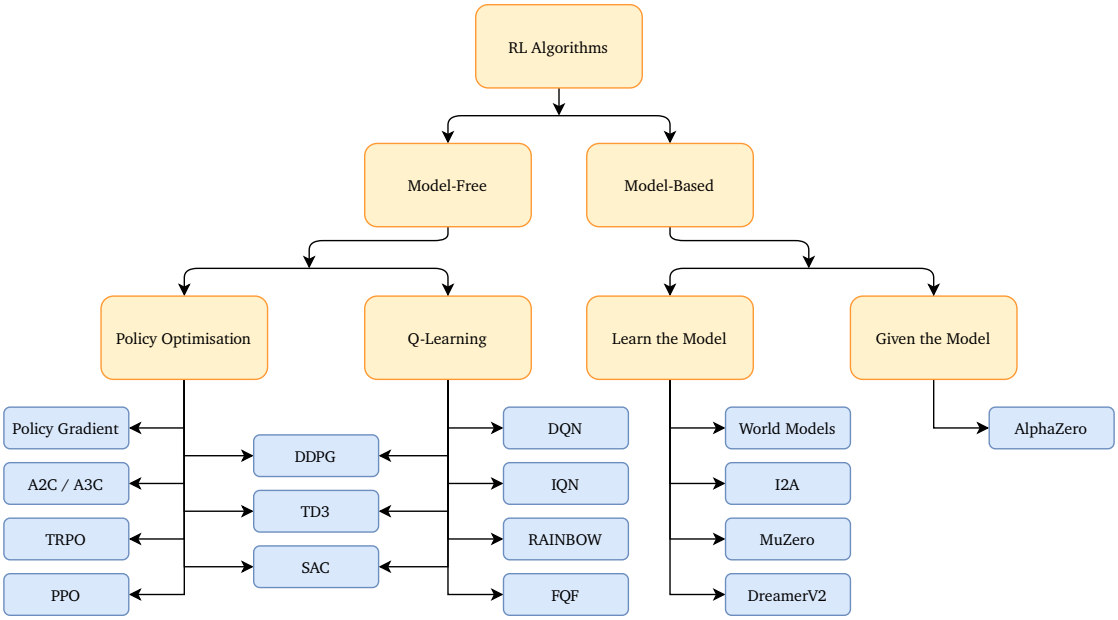


Figure 2: A non-exhaustive taxonomy of RL algorithms

DRL algorithms can be applied to complete robotic tasks. For example, mapless navigation, a motion planning task aims at navigating the robot close to the target and not collide with obstacles during the movement without obstacle map [2]. Building and updating global obstacle map is time-consuming and relies on precise dense laser scanner. With the help of emerging localisation methods, DRL can bring a relatively low-cost solution to this task

The remainder of the report is organised as follows: Project Description gives an overview of the project. Project Specifications lists what will be achieved. Methodology presents how to achieve the objectives. Project Plan lists tasks, milestones, and deliverables. Project Rationale and Industrial Relevance describles motivation from several perspectives. Literature Review discusses relevant previous works. Results reports the progress made so far. Finally, Conclusion concludes the report.

2 Project Description

The main objective of this project is to train an optimal policy for navigating a mobile robot to a desired target position with obstacle avoidance by deep reinforement learning in a simulated indoor environment full of obstacles. Four weeks before the end of the first semester, the robot is expected to be able to move to the target on a level ground with no obstacles. Before the end of semester one, it is expected that the robot can reach the target with obstacles. By the end of the project, the robot would be able to reach the target in unseen environment.

3 Project Specifications

- Literature review on mapless navigation with DRL.
- Implement one of the SOTA DRL algo (PPO/TD3/SAC) in one week.
- Train a policy in an empty world, four weeks before the end of the semester one.
- Set up an indoor simulation environment filled with obstacles for training in one week.
- Train a policy in the indoor environment, by the end of semester one.
- Set up another environment for evaluating the performance of the trained policy in unseen environment in one week.
- Train and evaluate the policy and select one which has the best performance in the evaluating environment, by the end of the project.
- Design the poster and prepare the demo for bench inspection.
- Document findings along the way in the final project report.

4 Methodology

The beginning of any development activity is research, so the first week is allocated to literature review. Use this step to find out what work has already been done by others and can be used directly.

~~Although~~ there are many libraries that provide implementation of required DRL algorithm~~, due to~~ the flexible nature of DRL~~, I prefer implementing it myself~~ in order to make it easier to add optimisation tricks later on.

Training the policy in an empty world ~~is~~ a sanity ~~test~~ to make sure ~~my~~ implementation of selected algorithm, simulator, virtual robot, reward function, and other components are compatible with each other.

The last step is standard deep learning model development practice which also fits into this DRL development project. The workflow ~~is basically train~~ the model on traning set, evaluate the performace of the trained model on ~~developing~~ set, repeat this loop several times ~~and~~ finally test the model on test set. However, to reduce the workload, the testing stage is removed from the aforementioned iterative development.

The programming language this project ~~will use~~ is Python. Python has a large ecosystem for machine learning~~,~~ with user-friendly frameworks such as PyTorch and JAX, and rich examples from reinforcement learning community ~~to follow~~.

The simulator will be Gazebo since it is integrated with Robotic Operating System (ROS) tightly, and ROS provides convenient approachs for data communication, which simplifies the robotics development greatly. In addition, ROS has great amount of robotics-related software packages include mobile robot models which can be imported into Gazebo easily.

5 Project Plan

The general plan of this project is presented in the form of a Gantt chart. See Appendix B. Tasks, milestones, and deliverables of this project are listed in Appendix C.

6 Project Rationale and Industrial Relevance

Traditional methods of motion planning build a map from sensor data and analyse this map to derive next actions. This type of methods is known as simultaneous localization and mapping (SLAM). Although this approach is reliable in large scale complex environments, the very high time and space resources required to maintain the maps make it too costly in small scale simple environments. DRL can perform motion planning in simple environment at low cost, though it does not guarantee reliability.

DRL-based motion planner is more competitive than SLAM for house robots that move only in ~~tidy~~ indoor environments. The computing systems of these robots usually have very little storage space and computing power with small battery capacity, which does not make them suitable for planning motions though mantaining an global obstacle map.

7 Literature Review

8 Results

DDPG and TD3 have been implemented and tested in Gym classic control and MuJoCo envs.

A simulated navigation training based on DDPG has been completed, but the performance was very poor due to frequent performance droping which is a limitation of DDPG. SAC, the successor to DDPG, ~~adds three tricks~~ to mitigate this limitation, and it is hoped that SAC will lead to better results ~~for~~ this project.

Algorithm 1 Soft Actor-Critic

Input: initial policy parameters θ, Q-function parameters ϕ_1, ϕ_2, empty replay buffer $\mathcal{D}$
Set target parameters equal to main parameters $\phi_{\text{targ},1} \leftarrow \phi_1$, $\phi_{\text{targ},2} \leftarrow \phi_2$
repeat
  Observe state s and select action $a \sim \pi_\theta(\cdot|s)$
  Execute a in the environment
  Observe next state s', reward r, and done signal d to indicate whether s' is terminal
  Store (s, a, r, s', d) in replay buffer $\mathcal{D}$
  If s' is terminal, reset environment state.
  if it's time to update **then**
    for j in range(however many updates) **do**
      Randomly sample a batch of transitions, $B = \{(s, a, r, s', d)\}$ from $\mathcal{D}$
      Compute targets for the Q functions:

$$y(r, s', d) = r + \gamma(1 - d)\left(\min_{i=1,2} Q_{\phi_{\text{targ},i}}(s', \tilde{a}') - \alpha \log \pi_\theta(\tilde{a}'|s')\right), \quad \tilde{a}' \sim \pi_\theta(\cdot|s')$$

      Update Q-functions by one step of gradient descent using

$$\nabla_{\phi_i} \frac{1}{|B|} \sum_{(s,a,r,s',d) \in B} \left(Q_{\phi_i}(s, a) - y(r, s', d)\right)^2 \qquad \text{for } i = 1, 2$$

      Update policy by one step of gradient ascent using

$$\nabla_\theta \frac{1}{|B|} \sum_{s \in B} \Big(\min_{i=1,2} Q_{\phi_i}(s, \tilde{a}_\theta(s)) - \alpha \log \pi_\theta\left(\tilde{a}_\theta(s)|\, s\right) \Big),$$

      where $\tilde{a}_\theta(s)$ is a sample from $\pi_\theta(\cdot|s)$ which is differentiable wrt θ via the reparametrization trick.
      Update target networks with

$$\phi_{\text{targ},i} \leftarrow \rho \phi_{\text{targ},i} + (1 - \rho)\phi_i \qquad \text{for } i = 1, 2$$

    end for
  end if
until convergence

Figure 3: SAC Pseudocode [1]

9 Conclusion

This report describes the background to the project including the basic theory of deep reinforcement learning, definition of mapless navigation, movtivation, and methodology. ~~Additionly,~~ this report specifies tasks, milestones, deliverables and timelines for each phase of the project. The last of the report provides a summary on progress to date.

References

[1] J. Achiam, 'Spinning Up in Deep Reinforcement Learning,' 2018.

[2] L. Tai, G. Paolo and M. Liu, 'Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation,' in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2017, pp. 31–36. DOI: 10.1109/IROS.2017.8202134.

Appendices

A Key specifications

Table 1: Specifications Verification Matrix

Specification	Verification
Train a policy in an empty world	Robot can reach the target position
Train a policy in an indoor environment filled with obstacles	Robot can reach the target position and not collide with obstacles
Trained policy can be used in unseen environment	Robot tries to move to the target while avoiding obstacle in an unexperienced environment

B Gantt chart

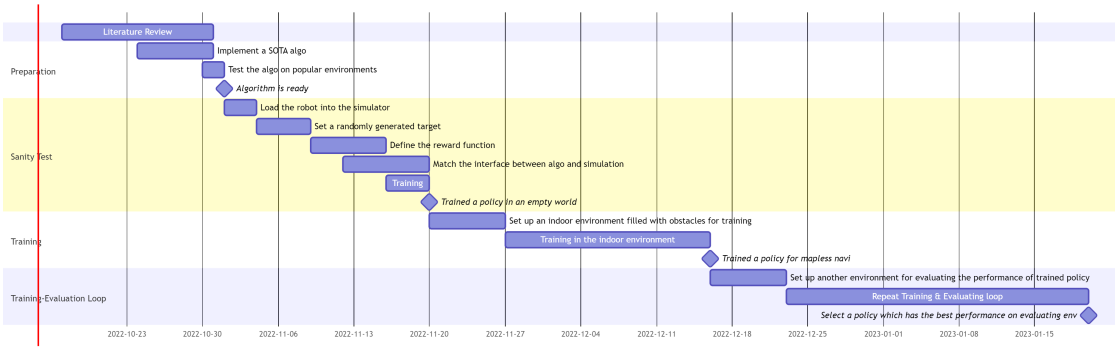


Figure 4: Gantt Chart

C List of work packages, milestones, and deliverables

- Literature review on mapless navigation with DRL.
- Implement one of the SOTA DRL algo (PPO/TD3/SAC) in one week.
- Train a policy in an empty world, four weeks before the end of the semester one.

- Set up an indoor simulation environment filled with obstacles for training in one week.
- Train a policy in the indoor environment, by the end of semester one.
- Set up another environment for evaluating the performance of the trained policy in unseen environment in one week.
- Train and evaluate the policy and select one which has the best performance in the evaluating environment, by the end of the project.
- Design the poster and prepare the demo for bench inspection.
- Document findings along the way in the final project report.

D Risk assessment form

UNIVERSITY OF LIVERPOOL

SINLGE USER BEng, MEng, MSc GROUP PROJECT RISK ASSESSMENT FORM - REPORT ONLY SIGNIFICANT HAZARDS

Unsafe working methods will lead to a reduction in your final project mark! ALL hardware work must be completed within the laboratory

Students are encouraged to come on site to perform their lab work but are advised that in some circumstances (Adriano, raspberry Pi and micro-controller boards which operate at <20V) equipment is allowed to be brought home. Students removing any other equipment from the lab needs to be authorised in writing by your supervisor - supervisors please confirm with HOD/safety team to confirm.

<table>
<tr><td colspan="2">NAME- Minghong Xu
Student ID Number- 201601082</td><td>LOCATION-
Final year Laboratory</td></tr>
<tr><td colspan="2">SCHOOL/DEPARTMENT: Electrical Engineering & Electronics
Undergraduate year of study: Third</td><td>BUILDING: Electrical Engineering and Electronics, A-Block</td></tr>
<tr><td colspan="3">TITLE OF PROJECT: Mapless Navigation with Deep Reinforcement Learning</td></tr>
<tr><td colspan="3">Description of Work: To train a policy function which enables a mobile robot to navigate to a target location without the need for a map by a state-of-the-art deep reinforcement learning algorithm (PPO/TD3/SAC) with optimisation extensions.</td></tr>
<tr><td rowspan="3">Select a category for this project:

Category
1 / 2 / 3</td><td colspan="2"><u>Category 1 – Projects based on specialist equipment:</u> Projects requiring equipment available in the electronics laboratories (such as power supplies, multimeters, oscilloscopes, etc.) or any other specialist equipment that requires specific health and safety considerations (such as drones, etc.) that students would not normally be allowed to take home.</td></tr>
<tr><td colspan="2"><u>Category 2 – Projects based on “home-friendly” equipment:</u> Projects requiring small pieces of equipment that do not require specific health and safety considerations and students can safely use at home (Raspberry Pi’s, Arduinos and other similar low-voltage boards with double insulated power supplies).</td></tr>
<tr><td colspan="2"><u>Category 3 – Projects based on software only:</u> Projects fully based on software that can be completed using only a computer, without requiring any other equipment.</td></tr>
</table>

If students are in an observation capacity only when experiment is being performed

- please state this on form as well as risk in being observers - i.e. possible distracting experimentalist,
- State risk if they could be injured in this respect and how. Significant risks only should be stated.
- Class of any laser is required

State voltage & current values of all power sources being used. Any power supplies that have the ability to generate current and voltages > 10mA AND >20V respectively can be regarded as potentially extremely hazardous: Voltage Current					
HAZARDS (Location, equipment and substances, activities)	**WHO CAN BE HARMED?**	**CURRENT CONTROLS**	**Likelihood (L) × Consequence (C) = RISK SCORE (R)**		
			L	**C**	**R**
On board regulators of Arty A7 provides maximum 5V 5A source	People nearby the board	5 A	1	1	1

- **For work using only Raspberry Pi and/or Arduino boards or other hardware connected via USB cable the main hazards are Display Screen Equipment (DSE) related, e.g. Repetitive Strain Injury, Carpal Tunnel Syndrome. L=1, C=1, R=1**

Training table - All boxes must be ticked in the following section to indicate either YES or NO.			
	N0	**YES**	**If you have ticked YES please follow the hyperlinks in the attached document, complete and return supplementary paperwork and/or implement and adhere to the guidance given.**
Use of tenon saw/hacksaw	✓		**Read Safe Operating** Procedure and other documentation on hand tools
Will work require the lifting of weights (>15kg)	✓		Manual Handling
Laser – If yes please input class of laser. Laser documents and hazard should be described on page 2 if laser is **NOT** class 1	✓		Please read all documents in the following link README : Laser: information and registration Guidance on the Safe Use of Lasers in Education & Research
Use gas cylinders or compressed gas?	✓		Gas Cylinder safety : Email local safety team to verify if training is required
Use hazardous Chemicals only? If stated on the form, description of hazard is required.	✓		COSHH - Use on-line EEE COSHH system to create COSHH risk assessment. Email local safety team to verify if training is required
Use voltages over 30V DC/AC If hazard has been previously described this	✓		Electrical Safety/Electricity – Includes reading the Sch. of EEE & CS dangers of electricity document
Use Power tools or rotating motors and machines	✓		SCR15-4 PUWER
Use Cryogenic Liquids/gases	✓		Cryogenic liquids and solids – Email local safety team to verify if training is required
Use Vacuum Systems and pressurised vessels	✓		Pressure systems : Email local safety team to verify if training is required
Use Radiation (UV, x-rays, microwaves)	✓		UV radiation (including links to local rules & safety advisor website)

LEVEL of Supervision?	A = Work May not be started without direct supervision
	B = Work may not start without Supervisor advice or approval
	C = No specific extra supervision requirements

Other relevant specific assessments (Local rules, Ethic approval forms)-

Disclaimer

- *The University of Liverpool ensures as far as is reasonably practical the health and safety of its staff and students.*
- *All equipment used by the students for their project must be safety tested and approved by the laboratory technicians before use. This includes but is not limited to, soldering irons, oscilloscopes, power supplies, probes and multimeters.*
- *Students **MUST NOT** undertake hazardous experimental/development work associated with their project outside of their designated laboratory space.*
- *ALL equipment that is used in the laboratory space & project MUST be purchased through the departments purchasing procedures.*
- *No equipment to be plugged into the mains supply unless circuit has been approved by technician or supervisor.*
- *Failure to abide by these conditions can result in the project receiving 0%.*
- ***Submission of this form implies acknowledgement by all the students named below.***

I can confirm that Hazards identified and precautions specified are appropriate for the task :-

Acknowledgement by Student 1 Name……Minghong…Xu…… Signature…… 徐铭鸿 ……Date……12…October…2022….

Academic supervisor Name……Murat Uney…… Signature…… ……Date……13 October 2022……

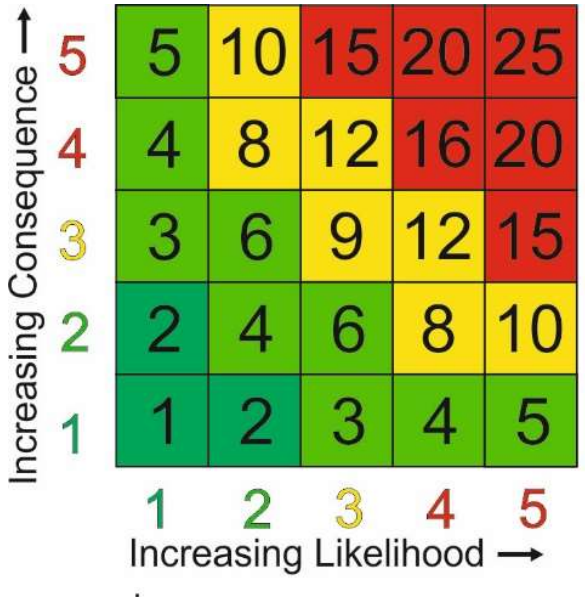
Common reasons for previously rejection of the form

- Project category was not stated on the assessment.
- Contradiction of hazards listed on page 2 compared those identified in training table. Users inserted description of hazards such as chemicals & live working but failed to insert yes in hazard table. Only hazardous chemicals should be described. Only significant hazards observed in experimental process should be described.
- Missing supervisor signature – risk assessment is invalid & students cannot enter the laboratory area
- Additional hazards noted in training table that are not described in hazard section. Lasers were described in training table required but hazard was not described in main assessment. Laser users should refer to risk assessment template document to identify how these should be described.

GUIDANCE TO COMPLETE THIS RISK ASSESSMENT FORM (LIKELIHOOD / CONSEQUENCE / RISK SCORE)

Likelihood			Consequence			Risk score	ACTION TO BE TAKEN
1	Very unlikely		1	Insignificant – no injury		1-2 NO ACTION	No action required but ensure controls are maintained and reviewed.
2	Unlikely		2	Minor – minor injuries needing first aid		3-9 MONITOR	Look to improve at next review of if there is a significant change
3	Fairly likely		3	Moderate – up to seven days absence		8-12 ACTION	Reduce risk if possible, within specified timescale
4	Likely		4	Major – more than seven days absence; major injury		15-25 STOP	Stop activity and immediate action
5	Very likely		5	Catastrophic – death; multiple serious injury			

- For work using only Raspberry Pi and/or Arduino boards (i.e. no other hardware connected using additional power supplies) the only hazards are Display Screen Equipment (DSE) related, e.g. Repetitive Strain Injury, Carpal Tunnel Syndrome. L=1, C=1, R=1



E Ethical approval questionnaire

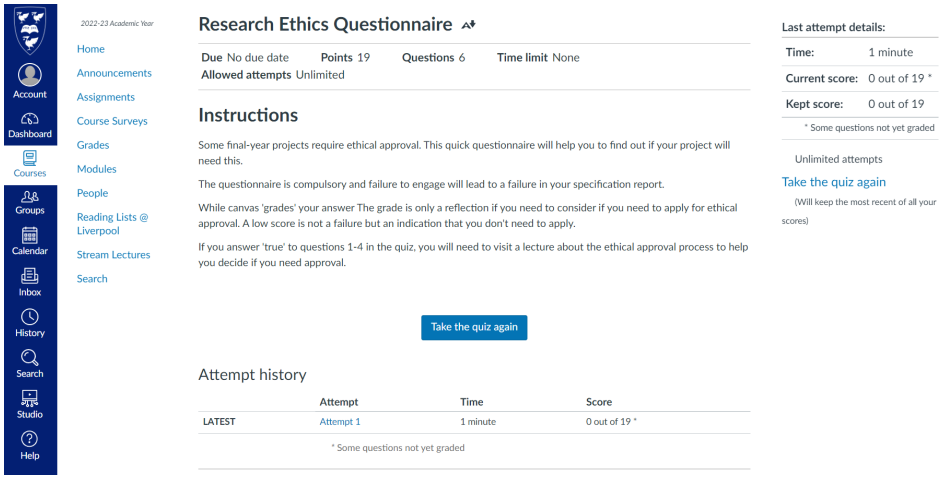


Figure 5: Screenshot of completion of Ethics Questionnaire