

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Symbolic Reinforcement Learning using Inductive Logic Programming

Author:
Kiyohito Kunii

Supervisor:
Professor Alessandra Russo

Submitted in partial fulfillment of the requirements for the MSc degree in MSc in
Computing Science of Imperial College London

June 2018

Contents

1	Introduction	2
2	Background	4
2.1	Logic Basics	4
2.1.1	Stable Model Semantics	5
2.1.2	Answer Set Programming (ASP)	5
2.2	Inductive Logic Programming (ILP)	6
2.2.1	Learning from Answer Sets (LAS)	7
2.2.2	Inductive Learning of Answer Set Programs (ILASP)	8
2.3	Reinforcement Learning	8
2.3.1	Markov Decision Process(MDP)	9
2.3.2	Temporal-Difference (TD) Learning	9
2.3.3	Q-Learning	10
2.3.4	Model-based vs model-free learning	10
3	Related Work	12
4	Project Overview	13
4.1	Proposed Architecture	13
4.1.1	ASP representation generation	13
4.1.2	Model generation and update using ILASP(RL)	14
4.2	Project outline	15
5	Ethics Checklist	16

Chapter 1

Introduction

There has been successful applications of deep reinforcement learning (DRL) a number of domains, such as games and robotics (TODO INSERT REFERENCES). DRL is considered to be a step towards artificial general intelligence, there are still a number of issues to overcome in this method. First, it requires a large amount of data for training the model, which requires a long time of learning process. Second, it is considered to be a black-box, meaning the decision making process is unknown to human user and therefore lacks explanation ability. Third there is no thoughts process to the decision making, which, as XX points out, is a fundamental to the artificial general intelligence. To tackle these problems, there are a number of different approaches the researchers have been experimenting. One of the methods is to incorporate symbolic representations into the system to achieve more data-efficient learning, which shows a promising results of this approach ([?]). In this paper, we explore the potential of inductive logic programming into reinforcement learning and attempt to solve the problems above. The advantages of symbolic reinforcement learning are, first of all, the decision making mechanism is understandable by humans rather than being black-box. Second, it resembles how human reason. There is some aspects of try-and-error in human learning, but they explore reasonings to efficiently learn the surrounding or situations. Also they effectively use previous experience (background knowledge) when encountered a similar situation. Finally, the recent advance of ILP research is remarkable and there are a number of new algorithm that effectively work in non-monotonic scenarios based on Answer Set Programmings (ASPs).

Recently there are a number of research that attempted to incorporate symbolic reasoning into DRL, but the combining inductive logic programming and reinforcement learning has not been explored. Because of the recent advancement of logic-based learning and deep reinforcement learning, combination of both approach would be a next explanation toward artificial general intelligence.

In this paper, our objectives is to explore this new research field and see how this combination of the two different learning methods could enhance the learning capability.

In this paper, I further explore incorporation of symbolic machine learning into reinforcement learning to achieve data-efficient learning using Inductive Learning of Answer Set Programs (ILASP), which is the state-of-art symbolic learning method

that can be applied to incomplete and more complex environment. This research is inspired by [?], but in this paper I explore symbolic representations process and include more learning aspect.

This background report is organised as follows: In Chapter 2, necessary background of logic, inductive logic programming and reinforcement learning are described for this paper. Chapter 3 discusses previous research in this particular approach and see the drawbacks of each approach. Chapter 4 shows the tentative architecture of our new approach, using Inductive Learning of Answer Set Programs (ILASP) to generate model of the environment. We also describe the outline of the project. Finally the ethics checklist is provided in Chapter 5.

Chapter 2

Background

This section introduces Logic Basics, Inductive Logic Programming (Section 2.2) and Reinforcement Learning (Section 2.3), which provide the foundations of this work.

2.1 Logic Basics

TODO ENTAILMENT

To compute this inference task, the syntax of the predicate logic program needs to be converted into Conjunctive Normal Form (CNF), or clausal theory, which consists of a conjunction of clauses, where a clause is disjunction of literals, and a literal can include positive literals and negative literals.

A literal is either an atom p or negation by failure, which is used to derive not p .

Example 2.1.1 (*Conjunctive Normal Form*)

The *Herbrand domain* (a.k.a *Herbrand universe*) of clause sets Th is the set of all ground terms that are constants and function symbols appeared in Th .

Example 2.1.2 (*Herbrand Domain*)

The *Herbrand base* of Th is the set of all ground predicates that are formed by predicate symbols in Th and terms in the *Herbrand domain*.

Example 2.1.3 (*Herbrand Base*)

The *grounding* of Th is the set of all clauses that are $c \in Th$ and variables are replaced by terms in the *Herbrand Domain*.

Example 2.1.4 (*Grounding*)

TODO Explain grounding in ASP context.

A *Horn clause* is a subset of CNF, which is a clause with at most one positive literal, where a *definite clause* (also called a *rule* or a *fact*) contains exactly one positive literal, and a *denial* (or a *constraint*) is a clause with no positive literals.

Example 2.1.5 (*Horn clause*)

Limitations of Inductive Logic Programming:

2.1.1 Stable Model Semantics

TODO Relationship from the previous Section

Definite Logic Program is a set of definite rule where:

a definite rule is of the form $h \leftarrow a_1, \dots, a_n$, where h, a_1, \dots, a_n are all atoms.

whereas

Normal Logic Program is a set of normal rule where

a normal rule is of the form $h \leftarrow a_1, \dots, a_n, \text{not } b_1, \dots, \text{not } b_n$ where h is the head of the rule, and $a_1, \dots, a_n, b_1, \dots, b_n$ are the body of the rule (both the head and body are all atoms).

The grounding of a normal logic program P can be obtained by replacing each rule in P with a ground instance of the rule, such that for each atom A in $\text{body}^+(R)$ (TODO EXPLAIN WHAT THIS IS), already occurs in the head of another ground rule.

RELATIONSHIP WITH PROLOG

The entire program needs to be grounded in order for ASP solver to work, and, unlike Prolog, each rule must be “safe”. A rule R is safe if every variable that occurs in the head of the rule occurs at least once in $\text{body}^+(R)$ (TODO check if this is correct).

An *Herbrand interpretation* of a set of definite clauses Th is a subset of the Herbrand base of Th , which is a set of ground atoms that are true in terms of interpretation.

Interpretation evaluate it to true Interpretation evaluate it to false

A *Herbrand Model* is a Herbrand interpretation if and only if a set Th of clauses is satisfiable.

Algorithm for computing Herbrand model is shown in

XXXX

A set of clauses Th is unsatisfiable if the algorithm terminates and does not return Herbrand model.

The *Least Herbrand Model* is an unique minimum Herbrand model if and only if none of its subsets is an Herbrand model (TODO MATHEMATICALLY SAY THIS)

For a normal program, there is usually no least Herbrand model. TODO Explain MORE

Stable Model of a normal logic program is XX [?]

Any stable model is a minimal Herbrand model, and stable sets is stable models. The stable models can be found by constructing the result of the program with respect to sets of atoms X (P^x in the following 2 steps

does not contain negation and a unique minimal Herbrand model.

Example 2.1.6 (*Horn clause*)

2.1.2 Answer Set Programming (ASP)

Answer set of normal logic program P is a stable model, and Answer Set Programming (ASP) is a normal logic program with extensions: constraints, choice rules and optimisation. ASP program consists of a set of rules, where each rule consists of an atom and literals. A literal is either an atom p or its *default negation* $\text{not } p$ (Negation as a failure).

Head and body

A Constraint is of the form

Theorem 1 $\leftarrow a_1, \dots, a_n, \text{not } b_1, \dots, \text{not } b_n$

which is to filtering any irrelevant answer sets. There are two types of constraints: soft and hard constraints. Soft constraint is XXX

Hard constraint is XXX

Choice rule can express possible outcomes given an action choice, which is of the form

Theorem 2 $lh_1, \dots, h_mu \leftarrow a_1, \dots, a_n, \text{not } b_1, \dots, \text{not } b_n$

where the head is called *aggregates*.

An answer set of ASP program is interpretations that make all the rules true.

Non-monotonicity.

ASP has true, false and unknown

for non-deterministic to describe transition possibilities.

optimisation statement is of the form.

Which is useful to order the answer sets in terms of preference.

Clingo is one of the programs to execute the ASP program and returns answer sets of the programs.

2.2 Inductive Logic Programming (ILP)

Inductive Logic Programming (ILP) is a subfield of research area aimed at the intersection between machine learning and logic programming [?]. The purpose of ILP is to inductively derive a hypothesis H that is a solution of a learning task, which covers all positive examples and any of negative examples.

ENTAILMENT

Satisfiable

$$B \cap H \models E \quad (2.1)$$

where E consists of positive examples (E^+) and negative examples (E^-)

One of the advantage of ILP over statistical machine learning is that the hypothesis the agent learnt can be easily understood by a human, making the decision more transparent rather than black box.

ILP in Answer set semantics

There are several ILP frameworks under the Answer set semantics.

Cautious and brave entailments can be defined in terms of induction in ILP [?].

Cautious Induction task is of the form $\langle B, E^+, E^- \rangle$ where:

B is the background knowledge

E^+ is a set of positive examples

E^- is a set of negative examples

$H \in \text{ILP}_{\text{cautious}} \langle B, E^+, E^- \rangle$ if and only if

there is at least one answer set A of $B \cup H$ such that:

for every answer set A of $B \cup H$: $\forall e \in E^+ : e \in A$

$\forall e \in E^- : e \notin A$

Example 2.2.1 (*Limitation of cautious induction*)

$1\{\text{situation}(P, \text{awake}), \text{situation}(P, \text{sleep})\}1 \text{ :- } \text{person}(P).$
 $\text{person}(\text{john}).$

In the example 2.2, neither of $\text{situation}(\text{john}, \text{awake})$ or $\text{situation}(\text{john}, \text{sleep})$ is false in all answer sets. In this example, it only return $\text{person}(\text{john})$.

Cautious entailment may be too restrict.

Similarly, Brave Induction task is of the form $\langle B, E^+, E^- \rangle$ where:

B is the background knowledge

E^+ is a set of positive examples

E^- is a set of negative examples

$H \in \text{ILP}_{\text{brave}} \langle B, E^+, E^- \rangle$ if and only if there is at least one answer set A of $B \cup H$ such that:

$\forall e \in E^+ : e \in A$

$\forall e \in E^- : e \notin A$

By contrast, Brave induction is that it cannot learn constraints as shown in the example 2.2.

Example 2.2.2 (*Limitation of brave induction*)

XXX

2.2.1 Learning from Answer Sets (LAS)

Cautious induction In order to overcome the limitations of cautious and brave induction, Learning from Answer Sets (LAS) was developed in [?] to facilitate more complex learning task.

Examples used in LAS is converted from $\langle E^+, E^- \rangle$ into

Partial Interpretations are of the form $\langle E^{\text{inc}}, E^{\text{exc}} \rangle$

A Herbrand Interpretations extends a partial interpretation if it include all of the inclusions and none of the exclusions.

2.2.2 Inductive Learning of Answer Set Programs (ILASP)

ILASP is an algorithm that is capable of solving LAS tasks

It is based on two fundamental concepts: positive solutions and violating solutions.

A hypothesis H is a positive solution if and only if 1. $H \subseteq S_M$

2. $\forall e^+ \in \exists A \in AS(B \cup H)$ subject to A extends e^+

A hypothesis H is a violating solution if and only if 1. $H \subseteq S_M$

2. $\forall e^+ \in E^+ \exists A \in AS(B \cup H)$ subject to A extends e^+

3. $\exists e^- \in E^- \exists A \in AS(B \cup H)$ subject to A extends e^-

ILP_{LAS} is positive solutions that are not violating solutions.

ILASP task containing a context-dependent example

Learning from Ordered Answer Sets

Weak constraints: useful for modelling the agent's preference

A Context-dependent Learning from Ordered Answer Sets

context-dependent partial interpretation (CDPI)

Context-dependent ordering example (CDOE)

Two advantages of adding context-dependent are it increases the efficiency of learning tasks, and more expressive structure of the background knowledge to particular examples.

2.3 Reinforcement Learning

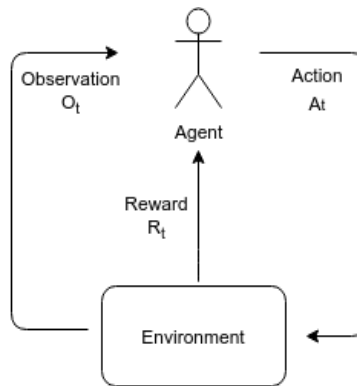


Figure 2.1: Agent and Environment

2.3.1 Markov Decision Process(MDP)

(Puterman, 1994) There is an agent who interacts with an environment (Figure 2.1). At each time step, an action taken by the agent affects the environment state and the reward (or penalty) it receives from the action outcome. (TODO Observations) When an agent must make a sequence of decisions, the sequential decision problem can be formalised using Markov decision process (MDP). MDPs formally represent a fully observable environment of an agent for reinforcement learning.

A MDP is of the form $\langle S, A, T_a, R_a, \gamma \rangle$ where:

- S is the set of finite states that is observable in the environment
- A is the set of finite actions taken by the agent
- $T_a(s, s')$ is a state transition in the form of probability matrix $\Pr(S_{t+1} = s' \mid s_t = s, a_t = a)$, which is the probability that action a in state s at time t will result in state s' at time $t+1$.
- R is a reward function $R_a(s, s') = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$, the expected immediate reward that action a in state s at time t will return
- γ is a discount factor $\gamma \in [0,1]$, which represents the preference of the agent for present rewards over future rewards

In MDPs, there is an element of delayed reward tradeoff between immediate rewards and its delayed reward

A state S_t is Markov if and only if $P[S_{t+1} \mid S_t] = P[S_{t+1} \mid S_1, \dots, S_t]$ therefore the probability of reaching S_{t+1} depends solely on S_t , which captures all the relevant information from earlier history.

A solution to the sequential decision problem is called a policy π , a sequence of actions that leads to a solution

The transition and reward functions are not necessary to be known to compute π .

An optimal policy π^* is the one that maximises the total rewards in the environment.

The total reward with a discount factor is

The existence of the discount factor can be justified as follows:

TODO $\mathbb{E}[R_{t+1} \mid S_t = s]$

Reinforcement learning is a method to get approximated optimal solution.

2.3.2 Temporal-Difference (TD) Learning

To solve MDP, one of the approaches is called Temporal-Difference (TD) Learning.

TD learns directly from episodes of experiences, which can be incomplete. TD does not require knowledge of MDP transitions and rewards (model-free) Sutton 1988

Update value

$$V(S_t) \leftarrow V(S_t) + \alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)) \quad (2.2)$$

where $R_{t+1} + \gamma V(S_{t+1})$ is the estimated return (a.k.a TD target)

$R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ is TD error.

TD updates the estimate by using the estimates of XXX (bootstrap).

The advantages of TD methods - does not require any model of an environment. - online learning

2.3.3 Q-Learning

model-free learning, where the agent will not rely on the models of the environment.

Q-learning approximate the $Q(a,s)$ from the samples of

the agent only knows about the possible states and actions but the transition state and reward probability functions are unknown.

Q function is the state-action pair

the optimal Q-function $Q^*(s,a)$ represents XXX for the agent to selection action a given that it is in state s .

Q-learning is off-policy TD learning defined in [Watkin 1989], which is of the form:

where α is the learning rate, γ is a discount rate between 0 and 1.

this equation is used to upate action-value function

Model free learning: derectly derive an optimal policy by interacting with the environment without the model

Model free can be done using Monte Carlo Policy evaluation

One way to solve the Bellman Optimality equation is Q-leraning

$U(s) = \max_a Q(s,a)$

a function $Q(S,A)$ which predicts the best action A in state S to maximise the total cumulative rewards.

The function is estimated by Q-learning, which repeately updates $Q(s,a)$ using the Bellman Equation.

TODO INSERT Q-learning ALGORITHM HERE

Epsilon greedy

2.3.4 Model-based vs model-free learning

A model M is a representation of an MDP where the state space and action space are assumed to be known. The model represents state transitions and rewards. In this case, the problem to be solved becomes a planning problem. A planning is for a series of actions to achive the agent's goal. Most of the reinforcement learning problems are model-free learning, where M is not unknown and the agnet learns to achieve the goal by the interactions with the environment. Thus the agent knows only possible states and actions, and the transition state and reward probability functions are unknown.

the performance of model-based RL is limited to optimal policy given the model M . Thus when the model is not a representation of the true MDP, the planning algorithms will not lead to the optimal policy, but suboptimal policy. The solution to this problem is either to use model-free approach when the model is inaccurate or incorporate undertainty of the model using XXX.

One algorithm which combine both aspects of model learning is called Dyna ([?]), which is shown in Figure 2.2.

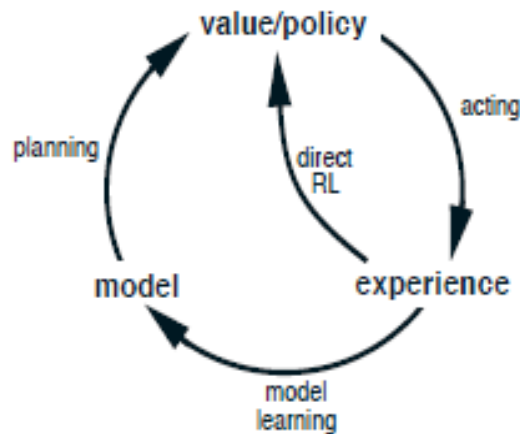


Figure 2.2: Relationships among learning, planning and acting ??

Dyna learns a model from real experience and use the model to generate simulated experience to update the evaluation functions.

This approach is more effective because the simulated experience is relatively easy to generate compared to having to have real experience, thus less iterations.

TODO ADD MATH

Chapter 3

Related Work

Transparency and interpretable capability of the model is another important aspect for machine learning applications.

XXX[Programmati...] developed a programmatically interpretable reinforcement learning which finds a policy that can be represented in a human-readable programming language.

Two most studied approach for using previous learning exprience is meta-learning and transfer learning

Artificial General Intelligence

Definision of AGI

The history of data-efficient learning What other people have done in this.

The advance of statistical machine learning methods, especially deeep reinforcement learning

AlphaGo, and AlphaGo Zero

Study of symbolic machine learning roots from

Baysian Optimisation

Symbolic Deep reinforcement learning

Some implementation: German paper

The paper was the application of symbolic representatiojns into a very simple game to demonstrate this proof of concept would actually work. By contract, there is active reserach in symbolic machine learning, which focuses on logic=based learning rather than statistical machine learning. For example, XX shows the agent can learn XXX from a noisy examples, with only a very few training examples.

Incorporation of logic into reinforcement learning dates back to the study of relational reinforcement learning,

More recently there has been a number of attempts to incorporate ASP into reinforcement learning.

There are a number of reseached conducted in applying DNN to symbolic reasoning. For example,

[From Gameplay to Symbolic Reasoning]

However, as XX points out, 1 pixel of the game could influence the decision making of the agent.

TODO Explain how symbolic learning works

Chapter 4

Project Overview

4.1 Proposed Architecture

The proposed tentative architecture is shown in Figure 4.1. The overall architecture resembles Sutton’s Dyna architecture ([?]), which combined both model-free learning and model-based learning.

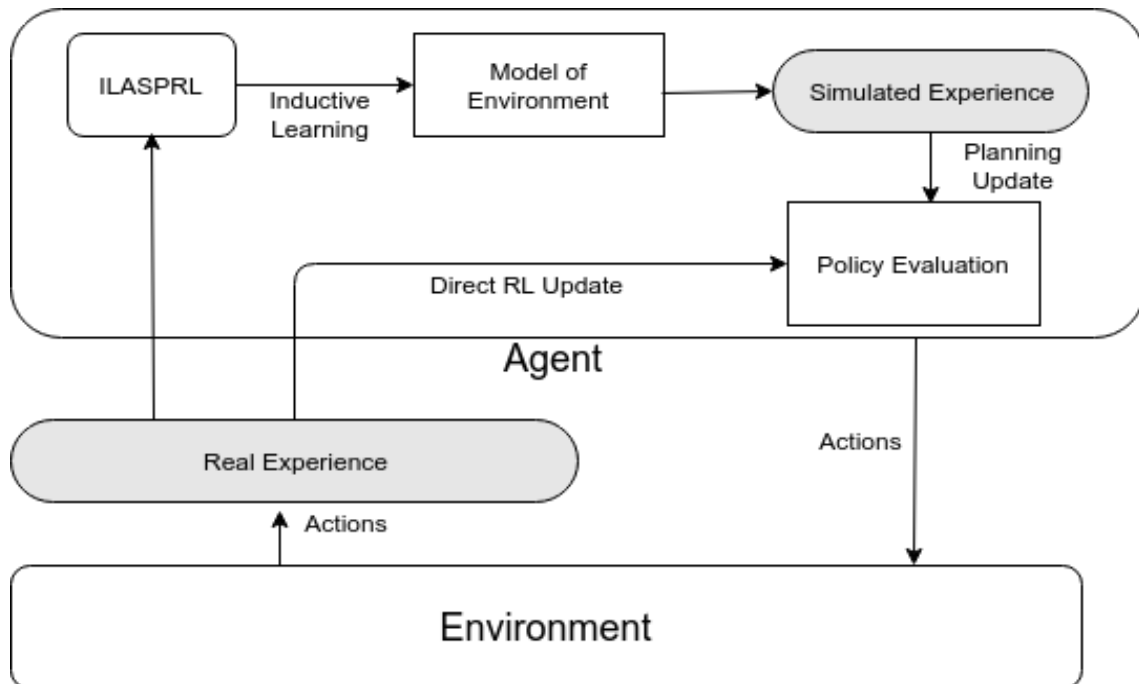


Figure 4.1: Proposed reinforcement learning architecture. ILASPRL learns to generate a model and updates based on the interaction with the environment, which is used to facilitate the policy evaluation.

4.1.1 ASP representation generation

The inputs from the experience needs to be converted in ASP form, which can be used to proceed the inductive learning in ILASP(RL). Observations from the inter-

action with the environment are state transition, reward and action of the agent, which can be directly converted using XXX. Example 4.1.1 is ASP input form that is feeded into ILASP(RL) phase.

Example 4.1.1 (*Examples for ILASPRL (input)*)

agent_before(S1, T1).

agent_after(S1, T1). reward(R). action(A).

4.1.2 Model generation and update using ILASP(RL)

Once the input is converted, the agent must learn the following definition of the model of the environment.

Example 4.1.2 (*Example of learning task (output)*)

valid_move(C0, C1):- previous(C0, C1).

previous(C2, C1):-

agent_at(C0, C1), adjacent(C0, C2), not obstacle(C0, C2), not enemy(C0, C2).

state_transition(S1, S2, T2):- agent_before(S1, T1), agent_after(S2, T2).

The background knowldge is empty, and each example contain different transition of the agent.

The model is updated as the agent explore more environment.

In addition, when the environment has been changed during the exploration, the agent creates a new model for this new environment.

In two different but similar scenarios, the model is generalised based on the two models, and the agent's model is more general which covers both games. Thus in theory, the agent should still be able to exploit the efficient learning from the model-based learning, facilitating the transfer learning.

Example: model refinement in a simple Grid world

Implement baseline performance (DQN, Q-learning, ASPRL)

Due to the complex environment of the chosen game, I will not implemnet the original DSRL, since the feature extractions from the pixel would only work in a very simple pixel environment,

Pipeline

The basic architecture follows the similar method in XXX, but I also add symbolic learning to these extracted symbolic features using ILASP.

Apply ILASP to the ASP, which involves development of the pipeline of ILASP in Python

Finally use Q-learning that allow

which measurement would you use? (grid word, something else? GVGAL games)

Summarise different types of knowledge representations (Objects ?? relationship?)

Common sense

Lastly, I will also test the capability of transfer learning for this new method.

4.2 Project outline

The next phase of the individual project is implementation and experiments. I will implement the proposed architecture using Python and ILASP (Clingo), and compare the performance

The platform I am planning to use to measure the performance of my approach will be GVG-AI Framework, which was created for the General Video Game AI Competition ¹, which provides game environments for an agent that should be able to play a wide variety of games without knowing which games are to be played. The underline language is the Video Game Definition Language (VGDL), which is a high-level description language for 2D video games providing a platform for computational intelligence research ([?]).

Especially, I will measure the following aspects of the new Algorithms

Learning efficiency

Transfer Learning capability

XXX

The proposed architecture is not finalised and will be reviewed regularly as I proceed the development of the architecture.

¹<http://www.gvgai.net/>

Chapter 5

Ethics Checklist

	Yes	No
Section 1: HUMAN EMBRYOS/FOETUSES		
Does your project involve Human Embryonic Stem Cells?		✓
Does your project involve the use of human embryos?		✓
Does your project involve the use of human foetal tissues / cells?		✓
Section 2: HUMANS		
Does your project involve human participants?		✓
Section 3: HUMAN CELLS / TISSUES		
Does your project involve human cells or tissues? (Other than from Human Embryos/Foetuses i.e. Section 1)?		✓
Section 4: PROTECTION OF PERSONAL DATA		
Does your project involve personal data collection and/or processing?		✓
Does it involve the collection and/or processing of sensitive personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?		✓
Does it involve processing of genetic information?		✓
Does it involve tracking or observation of participants? It should be noted that this issue is not limited to surveillance or localization data. It also applies to Wan data such as IP address, MACs, cookies etc.		✓
Does your project involve further processing of previously collected personal data (secondary use)? For example Does your project involve merging existing data sets?		✓
Section 5: ANIMALS		
Does your project involve animals?		✓
Section 6: DEVELOPING COUNTRIES		
Does your project involve developing countries?		✓

If your project involves low and/or lower-middle income countries, are any benefit-sharing actions planned?		✓
Could the situation in the country put the individuals taking part in the project at risk?		✓
Section 7: ENVIRONMENTAL PROTECTION AND SAFETY		
Does your project involve the use of elements that may cause harm to the environment, animals or plants?		✓
Does your project deal with endangered fauna and/or flora /protected areas?		✓
Does your project involve the use of elements that may cause harm to humans, including project staff?		✓
Does your project involve other harmful materials or equipment, e.g. high-powered laser systems?		✓
Section 8: DUAL USE		
Does your project have the potential for military applications?		✓
Does your project have an exclusive civilian application focus?		✓
Will your project use or produce goods or information that will require export licenses in accordance with legislation on dual use items?		✓
Does your project affect current standards in military ethics e.g., global ban on weapons of mass destruction, issues of proportionality, discrimination of combatants and accountability in drone and autonomous robotics developments, incendiary or laser weapons?		✓
Section 9: MISUSE		
Does your project have the potential for malevolent/criminal/terrorist abuse?		✓
Does your project involve information on/or the use of biological-, chemical-, nuclear/radiological-security sensitive materials and explosives, and means of their delivery?		✓
Does your project involve the development of technologies or the creation of information that could have severe negative impacts on human rights standards (e.g. privacy, stigmatization, discrimination), if misapplied?	✓	
Does your project have the potential for terrorist or criminal abuse e.g. infrastructural vulnerability studies, cybersecurity related project?		✓
Section 10: LEGAL ISSUES		
Will your project use or produce software for which there are copyright licensing implications?	✓	

Will your project use or produce goods or information for which there are data protection, or other legal implications?		✓
Section 11: OTHER ETHICS ISSUES		
Are there any other ethics issues that should be taken into consideration?		✓