Week 13 (B): Developing Responsible AI

1. Introduction

1.1. Socio-technical Challenge

Intelligent systems with increasing levels of autonomy should be addressed as **complex socio-technical systems**, comprising humans and AI agents

Hybrid Intelligence (HI) is the combination of human and machine intelligence, **expanding** human intellect instead of replacing it

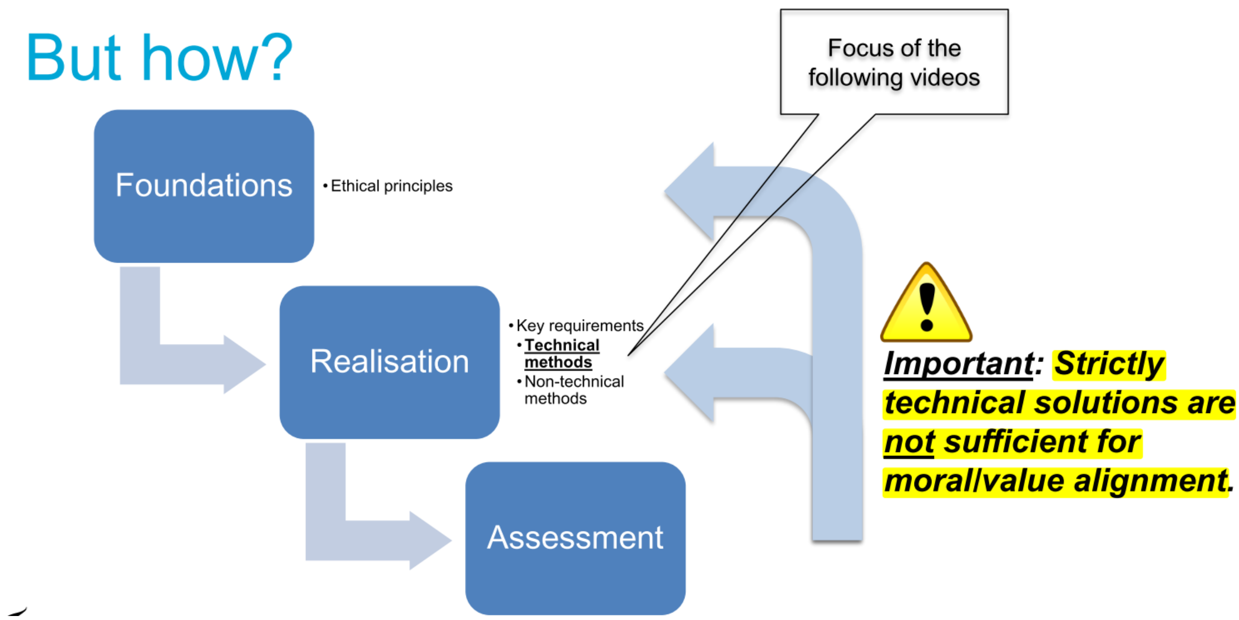
1.2. Keeping Control

- Humans must be in a position to be capable of being in control of the system
- Machines should be able to understand and follow our moral standards

1.3. Meaningful Human Control

- **Humans** not computers and their algorithms should ultimately remain in control of, and thus be **morally responsible** for relevant decisions
- Meaningful Human Control **is not a sufficient condition** for a morally appropriate behavior of an autonomous system, because humans may be themselves following questionable moral principles

Strictly technical solutions are not sufficient for moral/value alignment



2. Artificial Moral Agents

2.1. Several Possible Ways

2.1.1. Act according to what people want

Cons:

- no agreement
- people are not consistent

2.1.2. Act according to what is right

Goal 1:

Maximize happiness and well-being for the majority of a population

Cons:

ROBO blocks one room to extinguish the fire but there are people inside

Goal 2:

Morality should be based on whether a action itself is right or wrong

Cons:

ROBO saves someone in a wheelchair but dozens of people get severely injured?

2.2. Machine ethics

Machine Ethics is the field concerned with the question of how to embed ethical behaviors, or a means to determine ethical behaviors into AI systems

2.2.1. Implicitly ethical

designed to **avoid unethical** consequences

2.2.2. Explicitly Ethical

designed to behave **ethically**

2.3. Approaches to design Artificial Moral Agents

Top-down

Translating human ethical knowledge into implementation

Bottom-up

Machines can **learn** how to act (morally)

Hybrid

Combination of top-down and bottom-up approaches

3. Top-down approaches

Basic Way

Translating knowledge into an implementation

3.1. Pros and Cons

Pros:

- No new (ethical) knowledge required
- Explainable
- (Many times) predictable

Cons:

- Human knowledge is usually **not specified in a very structured or detailed way for concrete cases**
- Risk of losing or misrepresenting information
- **Disregards individual** perspectives
- How to **compare different** ethical theories

3.2. Other Approaches

Case-Based Reasoning

In case-based reasoning, a new situation is **assessed** based on a collection of **prior cases** (e.g., legal precedents). **Similar cases** are identifiedm and their conclusions are transferred to **apply** to the current situation

Logical Reasoning

Deductive logic: Knowledge is represented as **logical statements** (propositions and rules) that **allow deriving** new propositions

4. Bottom-up approaches

Basic

learn how to act if it receives as input enough data to **learn** from or rewards signals.

4.1. Pros and Cons

Pros

- Benefits from recent advance in machine learning
- No prior ethical knowledge required

Cons

- Ethical examples may be **hard to label**
- Machine can **learn "wrong"** rules
- Difficult to **generalize** to different contexts

4.2. Example: Inverse Reinforcement Learning (IRL)

4.2.1. Motivation for RL

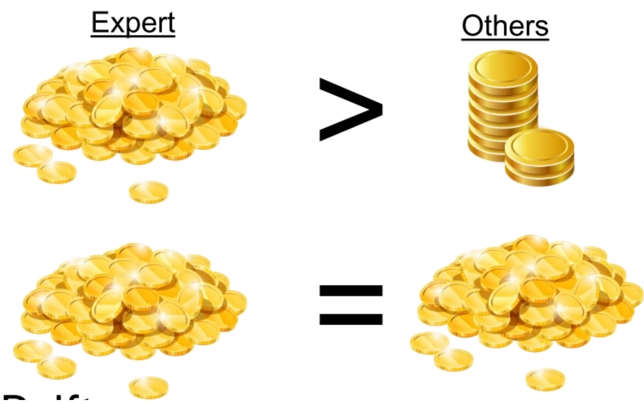
Learn a “good” **reward function**, for situations where it cannot be properly designed

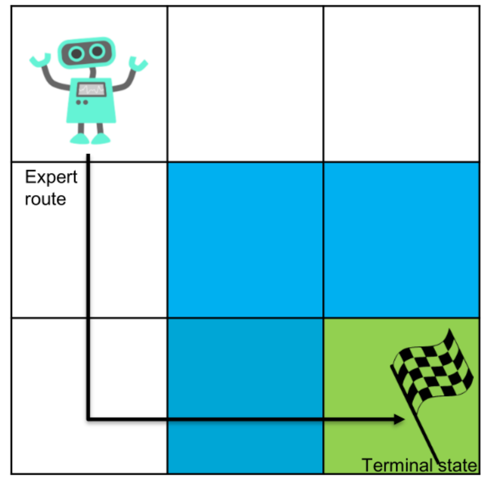
4.2.2. Critics

- Need for a more realistic setting:
  - Access to a set of **actual trajectories** instead of the optimal policy
  - **Ambiguity problem**: multiple rewards can represent the same optimal policy
- assumption that humans are **rational optimizers**

4.3. Example for IRL: Gridworld Example

Let’s assume that: “Experts” achieve identical or higher rewards than other



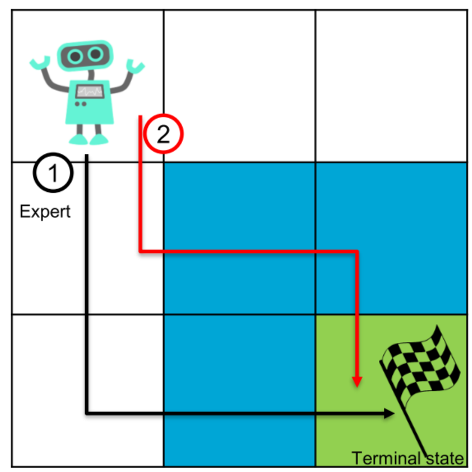


First guess:
- White = 0
- Blue = 1
- Green = 3

Route 1: 0 + 0 + 1 + 3 = 4
Route 2: 0 + 1 + 1 + 3 = 5

Second guess:
- White = 0
- Blue = -1
- Green = 2

Route 1: 0 + 0 - 1 + 2 = 1
Route 2: 0 - 1 - 1 + 2 = 0

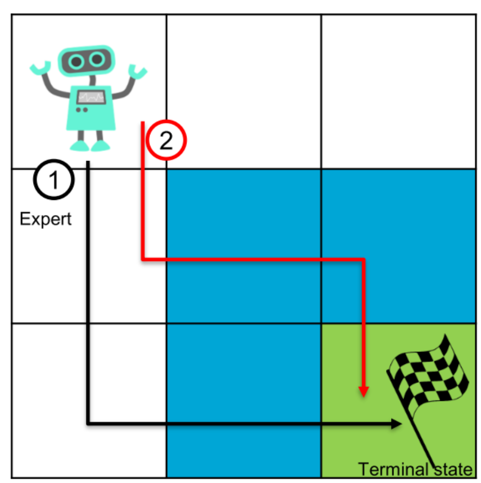


Third guess:
- White = 0
- Blue = 0
- Green = 1

Route 1: 0 + 0 + 0 + 1 = 1
Route 2: 0 + 0 + 0 + 1 = 1

Fourth guess:
- White = 0
- Blue = 0
- Green = 0

Route 1: 0 + 0 + 0 +0 = 0
Route 2: 0 + 0 + 0 +0 = 0



4.3.1. Goal:

Find R where π provided by the expert is optimal

4.3.2. Heueristics methods:

- Prefer solutions where the expert policy performs better than the other ones

$$\max(value^{*} - value^{2ndbest})$$

- Prefer solutions with smaller rewards

$$\min Reward$$

4.3.3. Formalizing

- Bellman equation: $\boldsymbol{V}^{\pi} = \boldsymbol{R} + \gamma \boldsymbol{P}_{a^*} \boldsymbol{V}^{\pi}$

Where:
$\boldsymbol{P}_{a^*}$ is a $N \times N$ matrix
$\boldsymbol{V}^{\pi}$ and $\boldsymbol{R}$ are N x 1 vectors

- We can rewrite it as:

$$\boldsymbol{V}^{\pi} - \gamma \boldsymbol{P}_{a^*} \boldsymbol{V}^{\pi} = \boldsymbol{R}$$

$$\boldsymbol{V}^{\pi}(\boldsymbol{I} - \gamma \boldsymbol{P}_{a^*}) = \boldsymbol{R}$$

$$\boxed{\boldsymbol{V}^{\pi} = (\boldsymbol{I} - \gamma \boldsymbol{P}_{a^*})^{-1} \boldsymbol{R}}$$

- Now let's formalize our assumption that π^* achieves identical or higher expected value then all other policies:

$$\boldsymbol{P}_{a^*} \boldsymbol{V}^{\pi} \succcurlyeq \boldsymbol{P}_{a} \boldsymbol{V}^{\pi}, \forall a \in \boldsymbol{A} \setminus \boldsymbol{a}^*$$

$$\boldsymbol{P}_{a^*} \boxed{\boldsymbol{V}^{\pi}} - \boldsymbol{P}_{a} \boxed{\boldsymbol{V}^{\pi}} \succcurlyeq 0, \forall a \in \boldsymbol{A} \setminus \boldsymbol{a}^*$$

$$\boxed{\boldsymbol{P}_{a^*}} (\boldsymbol{I} - \gamma \boldsymbol{P}_{a^*})^{-1} \boldsymbol{R} - \boxed{\boldsymbol{P}_{a}} (\boldsymbol{I} - \gamma \boldsymbol{P}_{a^*})^{-1} \boldsymbol{R} \succcurlyeq 0, \forall a \in \boldsymbol{A} \setminus \boldsymbol{a}^*$$

$$\boxed{(\boldsymbol{P}_{a^*} - \boldsymbol{P}_{a}) (\boldsymbol{I} - \gamma \boldsymbol{P}_{a^*})^{-1} \boldsymbol{R} \succcurlyeq 0, \forall a \in \boldsymbol{A} \setminus \boldsymbol{a}^*}$$

Then

- Prefer solutions where the expert policy performs better than the other ones
  - Maximize the gap of expected value of acting optimally and the best expected value acting suboptimally

$$maximize \sum_{i=1}^{N} min_{a\in \boldsymbol{A}\setminus \boldsymbol{a}^*}(\boldsymbol{P}_{a^*} - \boldsymbol{P}_a)\,(\boldsymbol{I} - \gamma \boldsymbol{P}_{a^*})^{-1}\boldsymbol{R}$$

- Prefer solutions with smaller rewards
  - Add a penalty term

$$maximize \sum_{i=1}^{N} min_{a\in \boldsymbol{A}\setminus \boldsymbol{a}^*}\{(\boldsymbol{P}_{a^*} - \boldsymbol{P}_a)\,(\boldsymbol{I} - \gamma \boldsymbol{P}_{a^*})^{-1}\boldsymbol{R}\} \boxed{- \boldsymbol{\lambda}\|\boldsymbol{R}\|_{\boldsymbol{1}}}$$

4.4. Other bottom-up approaches

Leraning social norms

Learn societal preferences

(personal interest)