Multi-agent Reinforcement Learning in Sequential social dilemmas

- **Machine-Machine** cooperation.
- In *Wolfpack* game, learning lone-wolf policy is easier than learning cooperative pack-hunting policy. This is because the former does not require actions to be conditioned on the presence of a partner within the capture radius.
- Greater network size leads to more cooperation.

In the **Gathering** game the situation is reversed. **Cooperative policies are easier** to learn since they need only be concerned with apples and may not depend on the rival player’s actions.

For Gathering, an **increase in network size** leads to an increase in the agent’s tendency to **defect**.

Source: [1]
• **Capacity** for more complex actions leads to more cooperative behaviour in the wolfpack, to less cooperation in the gathering game.

Increasing capacity does not automatically make the algorithm more cooperative.
• Shooting a beam might still be favourable, e.g. so that not both go for the same apple.
• Still need to improve cooperation.
• Possible ways to go at it:
  – Learn reward function for game.
  – Talk before you shoot. Communication is key.
Cooperating with Machines

• **Motivation**: Need algorithms to be able to cooperate, not just compete in special areas.

• **Goal**: AI algorithm cooperating with people/machines as good as humans cooperate (in arbitrary two-player repeated interactions).

• **Conditions** for successful algorithm: Generality, flexibility (associates), learning speed (human-machine)
• M-M and H-M cooperation.

• Standard ML algorithms could not bring players to cooperate effectively long-term.

• Idea: Introduce element of communication.
  – Helps to create shared representations.
  – **Cheap talk**: "Cheap talk refers to non-binding, unmediated, and costless communication"
- Cheap talk: Feedback and Planning.
- Difficulties: Some algorithm do not have easy understandable representations. But works with S++. 
- 19 possible sentences (different categories).

<table>
<thead>
<tr>
<th>Speech ID</th>
<th>Text</th>
<th>Speech ID</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Do as I say, or I’ll punish you.</td>
<td>10</td>
<td>We can both do better than this.</td>
</tr>
<tr>
<td>1</td>
<td>I accept your last proposal.</td>
<td>11</td>
<td>Curse you.</td>
</tr>
<tr>
<td>2</td>
<td>I don’t accept your proposal.</td>
<td>12</td>
<td>You betrayed me.</td>
</tr>
<tr>
<td>3</td>
<td>That’s not fair.</td>
<td>13</td>
<td>You will pay for this!</td>
</tr>
<tr>
<td>4</td>
<td>I don’t trust you.</td>
<td>14</td>
<td>In your face!</td>
</tr>
<tr>
<td>5</td>
<td>Excellent!</td>
<td>15</td>
<td>Let’s always play &lt;action pair&gt;.</td>
</tr>
<tr>
<td>6</td>
<td>Sweet. We are getting rich.</td>
<td>16</td>
<td>This round, let’s play &lt;action pair&gt;.</td>
</tr>
<tr>
<td>7</td>
<td>Give me another chance.</td>
<td>17</td>
<td>Don’t play &lt;action&gt;.</td>
</tr>
<tr>
<td>8</td>
<td>Okay. I forgive you.</td>
<td>18</td>
<td>Let’s alternate between &lt;action pair&gt; and &lt;action pair&gt;.</td>
</tr>
</tbody>
</table>
| 9         | I’m changing my strategy.                  |           | Source: [2]
• **Results:** Development of S# (Extension to S++)
  – S++ brings generality with it. Also fast convergence.
  – Communication via cheap talk is not the only, but on of the main features.
  – Info from communication reduces set of experts.
The chart illustrates the proportion of mutual cooperation in different experimental conditions: Human-Human, Human-S#, and S#-S#. The x-axis represents whether there was cheap talk (No or Yes), and the y-axis shows the proportion of mutual cooperation.

Source: [2]
• To forge mutually cooperative relationships, players must do two things: **Establish** cooperative behaviour and **maintain** it.
• Cheap talk helps with establishing (especially for humans)
• **Loyalty** is a reason for M-M pairs outperforming humans. Also **Honesty**.
Since verbal commitments by S# are derived from its intended behaviour, it does what it says.

Unlike "a sizeable portion" of the human participants.

Source: [2]
Over all games played, a human player had a positive net gain due to betrayals in just two interactions.
This graph does not necessarily imply that the AI is more evil than humans.

Maybe threats are just a more "effective" way to ensure cooperation.
Turing Test

% Thought to be Human

Partner  Human  S#

No  Yes
Cheap Talk?

Source: [2]
“The machine-learning algorithm learned to be loyal.”
(J. Crandall, Author)

This is open for discussion.

- **Big picture**: Added a new mechanism to the algorithm. Mimicking humans. Feedback by other player is used as part of the input.
CIRL - Cooperative inverse reinforcement learning³

“If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively [. . .] we had better be quite sure that the purpose put into the machine is the purpose which we really desire.” (Norbert Wiener, 1960)³

Value Alignment Principle: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.⁴

A **CIRL** problem is a cooperative partial information game:

- 2 agents, human and robot.
- Both rewarded according to the human’s reward function.
- But robot does not initially know what this is.

**Difference to IRL:**
- Observing agent (robot) is optimizing reward for the **human**.
- Acting Agent might act **suboptimal** to be better at explaining.
• Contribution: Optimal policy pair can be obtained by solving a **POMDP**.

• "Returning to Wiener’s warning, we believe that the best solution is not to put a specific purpose into the machine at all, but instead to design machines that provably converge to the right purpose as they go along."
Problem with Value Alignment: What Do We Want?

Understanding what “we” want is a big challenge for AI research.

– Difficult to *encode* human values in a programming language.
– Humanity does not agree on *common values*.
– And the parts we do agree on change with time.

Are human values the *best values* there can be?
“Friendly” AGI via Human Emotion: the Vital Link

- Consider trade-off situation, ethical dilemma.

- How does a busy AGI even be\textit{come aware} that a situation calls for an ethical action or decision?

- Control by human intercession \textit{not feasible}.

- \textbf{Recognition} of a problem as first step.

\[5\]

• **Decision making**: Limited computational capacity (always information overflow) → distillation/filtering of info (just like humans).

• Make decision based on: memory, pattern recognition, prediction, evaluation.
  → AGI confronted with **same problems as humans**.
• **Difference Engine**: Expected an perceived states. Situation with great discrepancy between expected and perceived will receive attention. ➔ **Homeostasis**

• Valuation of this disparity? Emotions and Needs.
• Include humans in the needs of the AGI.
• Needs of AGI? minimally physical, social needs, data security?
• Needs distinction between self and others.
• Who is me? Make "We" and "Me" inseparable, so that it includes the human team.
It is critical than humans are innate members of the AGI ingroup.
Arguments against Linking Human Emotion and AGI in Meta-Beings: Privacy, Freedom, Individuality; who dominates, whose needs dominate?

"Multi-individual homo communicatus, joined through our technologies"\textsuperscript{5}

Communication might be a key in linking.
Take away

• Role of communication as mechanism in AI algorithms.
• Idea of teaching the AI human values via IRL.

Mechanisms and algorithms can be used to introduce concepts of cooperation into AI.
• Might not be sufficient.
• Maybe resolve the problem by going from "it"/"us" to just "us".

Meta-beings
Sources


and https://futureoflife.org/ai-principles/

• james barrat 2013 ai
• harming humans. asimow
Whitaker Paper: Modeling of Donation games

• Social brain hypothesis

• Social Heuristics Hypothesis:
behaviours that support success in regular social interactions become intuitive and automatic (type-1, intuitive), unless they are moderated by reflective type-2 (cognitive) processes that represent learning to update a type-1 heuristic.
• Using "social comparison", reputation.
• Avoid free riders.
• results showed that evolution favours the strategy to donate to those who are at least as reputable as oneself

Big picture: Introduced a score