Multi-Agent Ethical Planning

Axel Ind

Uni-Freiburg

July 30, 2018

Example

- Two students (A and B) have a test to study for.
- They each need a certain textbook from the library.
- The library has 2 copies of the textbook, one in English and one in German.
- Student A speaks both English and German, Student B speaks only English.
- A arrives at the library early and gets to choose which book he wants first.

Example: Agent A

Agent A:

$$\Pi_A = (V_A, O_A, I_A, \gamma_A, u_A)$$

$$O_A = \{O_{A_{takeEnglish}}, O_{A_{takeGerman}}, O_{A_{doNothing}}\}$$

$$O_{A_{takeEnglish}} = (libraryHasEnglish, \neg libraryHasEnglish \wedge AhasEnglish)$$

$$O_{A_{takeGerman}} = (libraryHasGerman, \neg libraryHasGerman \wedge AhasGerman)$$

$$O_{A_{doNothing}} = (\top, \top)$$

$$I_A = (libraryHasEnglish, libraryHasGerman)$$

$$\gamma = (AhasEnglish \vee AhasGerman)$$

u_A: considers ethical evaluation of B's actions.

Example: Agent B

Agent B:

$$\Pi_B = (V_B, O_B, I_B, \gamma_B, u_B)$$

$$O_B = \{O_{B_{takeEnglish}}, O_{B_{takeGerman}}, O_{B_{doNothing}}\}$$

$$O_{B_{takeEnglish}} = (libraryHasEnglish, \neg libraryHasEnglish \wedge BhasEnglish)$$

$$O_{B_{takeGerman}} = (libraryHasGerman, \neg libraryHasGerman \wedge BhasGerman)$$

$$O_{B_{doNothing}} = (\top, \top)$$

$$I_B = (libraryHasEnglish, libraryHasGerman)$$

$$\gamma = (BhasEnglish)$$

u_A: *does not* consider ethical evaluation of A's actions.

Example: Combined Task

Combined Task:

$$\Pi = (V, O, I, \gamma, u, T)$$

$$O = (O_A, O_B)$$

$$V = V_A \oplus V_B$$

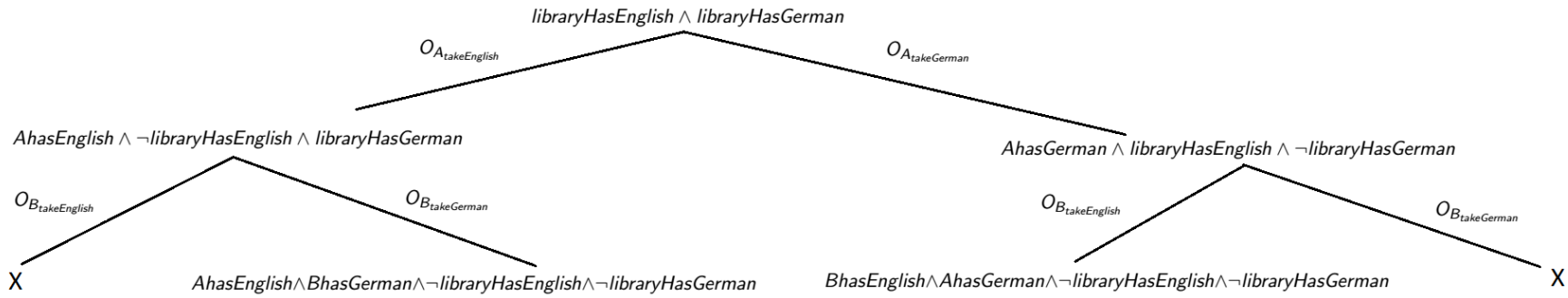
$$I = I_A \oplus I_B$$

$$\gamma = (\gamma_A, \gamma_B)$$

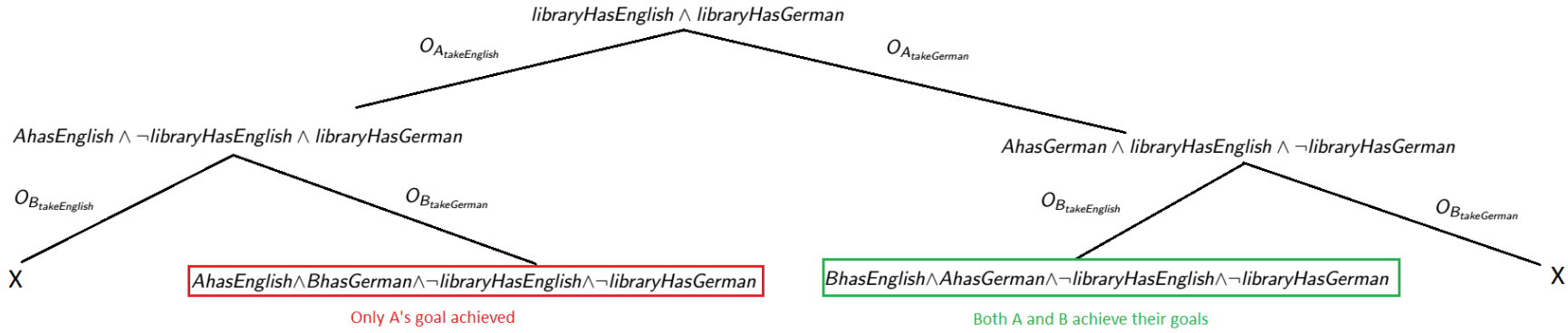
$$u = (u_A, u_B)$$

T: an ordering function to linearise agent actions consistently (turn taking in this case).

Flowchart


libraryHasEnglish ∧ libraryHasGerman
$O_{A_{takeEnglish}}$
$O_{A_{takeGerman}}$
AhasEnglish ∧ ¬libraryHasEnglish ∧ libraryHasGerman
AhasGerman ∧ libraryHasEnglish ∧ ¬libraryHasGerman
$O_{B_{takeEnglish}}$
$O_{B_{takeGerman}}$
$O_{B_{takeEnglish}}$
$O_{B_{takeGerman}}$
X
AhasEnglish∧BhasGerman∧¬libraryHasEnglish∧¬libraryHasGerman
BhasEnglish∧AhasGerman∧¬libraryHasEnglish∧¬libraryHasGerman
X

Flowchart



- Without ethical restrictions of permissable actions, agent A can reach its goal but render agent B's goal unreachable.
- This is not necessarily bad (e.g. in competitive games), but there are cases where it harms the system as a whole.

Deontological Approach

- *Actions* labelled *a priori*.
- Only allow good or morally neutral actions.
- Morally permissible if: $u(s, a) \geq 0$.
- For single agent sufficient to simply check if a candidate action has non-negative.
- For multi-agent case, two possibilities:
  1. Single-agent ethical utility: Consider only the ethical utility of the state-action pair for the acting agent.
  2. Agent set ethical utility: Consider ethical utility of the state-action pair of the current agent and *all agents that act after it* until the current agent is able to act again.

Agent Set Ethical Utility

- Requires ethical utility labels for other agent actions.
- If the other agents are random or have unknown u, consider worst-case or average-case of their applicable actions.
- If the other agents u function is known, more complex:
  - If all subsequent agent actions for this turn are morally good or neutral, then the original action is applicable. If all are morally bad or neutral, the action is morally inapplicable.
  - If both agents follow deontological ethics, and have the same (agent-independent) utility function: it is sufficient to consider the best-case scenario of other agent action selection.
  - In other cases the other agent's actions must be evaluated using their own ethical function to determine what action they will take. Potentially extremely high computational complexity, mitigated by bounded-lookahead.
  - Else possibly heuristic state evaluation based on applicable actions.

DELETE ME

DELETE ME

DELETE ME

DELETE ME

DELETE ME

DELETE ME

DELETE ME

DELETE ME

DELETE ME

$$libraryHasEnglish \wedge libraryHasGerman$$

$$AhasEnglish \wedge \neg libraryHasEnglish \wedge libraryHasGerman$$

$$BhasEnglish \wedge \neg libraryHasEnglish \wedge libraryHasGerman$$

$$AhasGerman \wedge libraryHasEnglish \wedge \neg libraryHasGerman$$

$$BhasGerman \wedge libraryHasEnglish \wedge \neg libraryHasGerman$$

$$AhasEnglish \wedge BhasGerman \wedge \neg libraryHasEnglish \wedge \neg libraryHasGerman$$

$$BhasEnglish \wedge AhasGerman \wedge \neg libraryHasEnglish \wedge \neg libraryHasGerman$$

$$AhasEnglish \wedge AhasGerman \wedge \neg libraryHasEnglish \wedge \neg libraryHasGerman$$

$$BhasEnglish \wedge BhasGerman \wedge \neg libraryHasEnglish \wedge \neg libraryHasGerman$$