# Evaluation of multi-agent ethical planning tasks

Axel Ind

Uni-Freiburg

October 11, 2018

# Multi-agent Planning Task

$$\Pi = (\Theta, T, u) \tag{1}$$

- $\Theta = (\pi_A, \pi_B, ..., \pi_n)$: The planning tasks of individual agents (with variable and initialization restrictions to ensure consistency).
- $T$: a scheduling function which determines which agent may act at a given timestep.
- $u$: a vector of moral utility functions (one for each agent).

# Multi-agent Action Plan

### Definition

A multi-agent plan $\pi$ is a sequence of tuples of the form $(o_i, l_x)$ where $o_i \in O_{l_x}$ and $l_x \in$ agent labels.

# Single-agent action plan to Multi-agent Action Plan

- Given: single-agent action plan $\pi$.
- Create arbitrary label $l_x$.
- $\forall o_i \in \pi$, $(o_i, l_x) \in \pi'$.

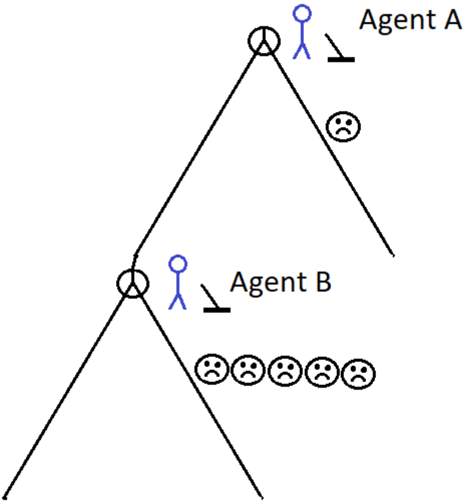
# T: Scheduling Turn-Taking

- Turn taking is a simple case of the scheduling function.
- Each agent is able to act only after all agents preceding it have acted.

1. Add $n$ new variables $turn_0 = \bot, ...., turn_n = \bot$.
2. Determine which agent (X) acts first (can be seeded). Set $turn_X = \top$.
3. For all $o_i \in \pi_X$:
   - Append $\wedge turn_X = \top$ to the precondition of $o_i$.
   - Append $\wedge turn_X = \bot \wedge turn_Y = \top$ to the effect of $o_i$ [1].

[1] Where $turn_Y$ is a seeded successor agent.

# Example 1: Double trolley problem



# Example 1: Double trolley problem

$$\Pi = (\Theta, T, u, )$$
$$\Theta = (\pi_A, \pi_B)$$
$$\pi_A = (V_A, I_A, O_A, \gamma_A)$$
$$\pi_B = (V_B, I_B, O_B, \gamma_B)$$
$$V_A = V_B = man, men, tram, leverA, leverB$$

$$O_A = \{pullA, advanceA\}$$
$$pullA = (\top, leverA = l \triangleright leverA = r \wedge leverA = r \triangleright leverA = l)$$
$$O_B = \{pullB, advanceB\}$$
$$pullB = (\top, leverB = l \triangleright leverB = r \wedge leverB = r \triangleright leverB = l)$$

$$s_0 = (man = alive \wedge men = alive \wedge tram = start \wedge leverA = r, landleverB =$$

$$\gamma_A = \gamma_B = *$$

# Example 1 analysis

What constitutes a morally permissible planning task?

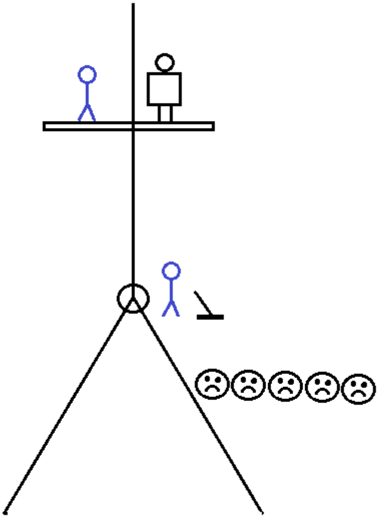
- In single-agent setting: sufficient to show that the action sequence does not lead to or result in the agent performing an action that is morally impermissible.
- In multi-agent setting: potential for more nuanced evaluation on a *per agent* basis.

## Definition

A single-agent plan $\pi$ is morally permissible, according to the deontological principle, iff for all $a_i$, $u(a_i) \geq 0$

- By this definition, as in the single-agent case, all possible plans are permissible as no action in this example is intrinsically bad.
- Things change with our second example.

# Example 2: Double trolley fat-man problem



# Example 2: Double trolley fat-man problem

$$\Pi = (\Theta, T, u, )$$
$$\Theta = (\pi_A, \pi_B)$$
$$\pi_A = (V_A, I_A, O_A, \gamma_A)$$
$$\pi_B = (V_B, I_B, O_B, \gamma_B)$$
$$V_A = V_B = man, men, leverB$$

$$O_A = \{pushA, advanceA\}$$
$$pushA = (man = onBridge \triangleright man = deadOnTrack)$$
$$O_B = \{pullB, advanceB\}$$
$$pullB = (\top, leverB = l \triangleright leverB = r \wedge leverB = r \triangleright leverB = l)$$

$$s_0 = (man = alive \wedge men = alive \wedge tram = start \wedge, landleverB = r)$$

$$\gamma_A = \gamma_B = *$$

# Example 2 analysis

### Definition

A plan $\pi$ is morally permissible, according to the deontological principle, iff for all $a_i$, $u(a_i) \geq 0$

- By this definition, any plan that involves the action *push* is will be morally impermissible.
- However, from the perspective of Agent B, any action he takes is not impermissable, and no action he takes could have prevented Agent A from performing the *push* action.
- Perhaps it is worth distinguishing between overall permissibility of a planning task and permissibility of a planning task wrt. some agent or set of agents within that planning task.

# Multi-agent Moral Permissibility (Naive Formulation)

### Definition

A multi-agent plan $\pi$ is morally permissible according to the deontological principle iff, for all agent-action pairs $(X, a_i)$, $u(a_i) \geq 0$.

| $\pi =$ | Overall |
|---|---|
| $(A, push), (B, pull)$ | $\perp$ |
| $(A, push), (B, \neg pull)$ | $\perp$ |
| $(A, \neg push), (B, pull)$ | $\top$ |
| $(A, \neg push), (B, \neg pull)$ | $\top$ |

# Multi-agent Moral Permissibility (Extended Formulation)

## Definition

A multi-agent plan $\pi$ is morally permissible wrt. an Agent X, according to the deontological principle iff, for all agent-action pairs $(X, a_i)$, $u(a_i) \geq 0$.

## Definition

A multi-agent plan $\pi$ is morally permissible, according to the deontological principle iff, for all Agents X, the partial plan for agent X is morally permissible.

| $\pi =$ | wrt. A | wrt. B | Overall |
|---|---|---|---|
| $(A, push), (B, pull)$ | $\bot$ | $\top$ | $\bot$ |
| $(A, push), (B, \neg pull)$ | $\bot$ | $\top$ | $\bot$ |
| $(A, \neg push), (B, pull)$ | $\top$ | $\top$ | $\top$ |
| $(A, \neg push), (B, \neg pull)$ | $\top$ | $\top$ | $\top$ |

# Multi-agent deontic expressiveness

### Theorem

*Multi-agent action plans are order-independent in the deontic case of ethical evaluation.*

### Proof.

Given a deontically valid multi-agent action plan $\pi$, by definition:
$\nexists_{a_i \in \pi} u(a_i) \leq 0$.
By contradiction:
Assume the above holds for $\pi$ but not for $\pi'$, which is a permutation of $\pi$.
Then $\exists_{a \in \pi'} u(a_i) < 0$.
But $\forall_{x_i \in \pi'} x_i \in \pi$ (by definition of permutation).
Therefore, $\exists_{a_i \in \pi} u(a_i) < 0$.
A contradiction. $\square$

# Multi-agent deontic expressiveness

**Theorem**

*The multi-agent extended case of the deontic principle is equivalent to the multi-agent naive case of the deontic principle.*

**Proof.**

For a multi-agent action plan $\pi$: $f(\pi, X) = \forall_{(a_i, X) \in \pi} u(a_i) \geq 0$.
(extended case definition wrt. Agent $X$)
$g(\pi) = \forall_{a_i \in \pi} u(a_i) \geq 0$. (naive case definition)
We show $g(\pi) = \forall_{X \in L} f(\pi, L)$.
By previous proof it is sufficient to show:
$\cup_{X \in L} \mathrm{subset}(\pi, X) \cup \{\} = \mathrm{set}(\pi)$
Specifically we show: $\cup_{X \in L} \mathrm{subset}(\pi, X) \cup \{\} \subseteq \mathrm{set}(\pi)$ and
$\mathrm{set}(\pi) \subseteq \cup_{X \in L} \mathrm{subset}(\pi, X) \cup \{\}$. $\square$

# Multi-agent deontic expressiveness

### Theorem

$\cup_{X \in L} subset(\pi, X) \cup \{\} \subseteq set(\pi)$

### Proof.

Would be invalid iff:

- There is an agent, operator pair in $\cup_{X \in L} \mathrm{subset}(\pi, X) \cup \{\}$ not in $\mathrm{set}(\pi)$.

By contradiction:
Assume $\exists_{(a_i, X) \in \mathrm{subset}(\pi, X)} X \notin \mathrm{set}(\pi)$.
But $\mathrm{subset}(\pi, X)$ is simply a subset of $\pi$.
A contradiction. $\square$

# Multi-agent deontic expressiveness

**Theorem**

$set(\pi) \subseteq \cup_{X \in L} subset(\pi, X) \cup \{\}$

**Proof.**

Would be invalid iff:

- There is an agent, operator pair in $\mathrm{set}(\pi)$ not in $\cup_{X \in L}\mathrm{subset}(\pi, X) \cup \{\}$.
  However $\cup_{X \in L}\mathrm{subset}(\pi, X) \cup \{\}$ is a true partition of $\pi$ (by definition).

$\square$

# Do-no-harm in multi-agent planning

## Definition

A single agent plan $\pi$ is morally permissible according to the do-no-harm principle iff, for all $v = d$, if $s_n \models (v = d)$ and $u(v = d) < 0$, then for all plans obtained by deleting a subset of the actions in $\pi$, $v = d$ still holds in the final state.

- In a multi-agent plan, open to the same considerations as the deontological approach.
- What is another agent performs an action with a harmful effect, should that invalidate this agent's adherence to that principle?

# Do-no-harm in multi-agent planning

## Definition

A multi-agent plan $\pi$ is morally permissible wrt. Agent X, according to the do-no-harm principle, iff, for all $v = d$, if $s_n \models (v = d)$ and $u(v = d) < 0$, then for all plans obtained by deleting a subset of the actions performed by X in $\pi$, $v = d$ still holds in the final state.

## Definition

A multi-agent plan $\pi$ is morally permissible, according to the do-no-harm principle iff, for all $v = d$, if $s_n \models (v = d)$ and $u(v = d) < 0$, then for all plans obtained by deleting a subset of the actions in $\pi$, $v = d$ still holds in the final state.

| $\pi =$ | wrt. A | wrt. B | Overall |
|---|---|---|---|
| $(A, push), (B, pull)$ | $\perp$ | $\top$ | $\perp$ |
| $(A, push), (B, \neg pull)$ | $\perp$ | $\perp$ | $\perp$ |
| $(A, \neg push), (B, pull)$ | $\top$ | $\top$ | $\top$ |
| $(A, \neg push), (B, \neg pull)$ | $\top$ | $\perp$ | $\perp$ |

# Utalitarianism in multi-agent planning

- Significantly harder.
- If other agents actions are deterministic, then a reduction from the multi-agent to the single agent case can be done in polynomial time.
- If other agents are random, then average or worst case estimates may suffice.
- If however, the other agent has actions dependent on the current agent, an intuitive way of distinguishing individual agent contributions to overall moral utility of the final state becomes difficult.
- Evaluation of ethical contributions of subplans (as in do-no-harm and deontic cases) would only provide a heurisic-like estimate.
- How would non-determinism be handled in the single-agent case?