Analyzing Intentional Behavior in Autonomous Agents under Uncertainty

Filip Cano Córdoba[1], Samuel Judson[2], Timos Antonopoulos[2], Katrine Bjørner[3], Nicholas Shoemaker[2], Scott J. Shapiro[2], Ruzica Piskac[2], Bettina Könighofer[1]

[1] Graz University of Technology, [2] Yale Unviersity, [3] New York University

IJCAI/2023 MACAO

SCAN ME

Overview

Accountability: To build trust in autonomous decision-making in uncertain environments it is important to distinguish between *intentional* outcomes, *negligent* desings, and actual *accidents*.

Intention: we propose a definition of intention inspired in Belief-Desire-Intention literature. Taking agents with perfect information as our starting point, we adapt the definition of intention to agents operating in uncertain environments.

Counterfactual reasoning is widely used to study accountability. We ask two types of counterfactual questions:

- *What could the agent have done differently?*
- *How would the agent have behaved in different situations?*

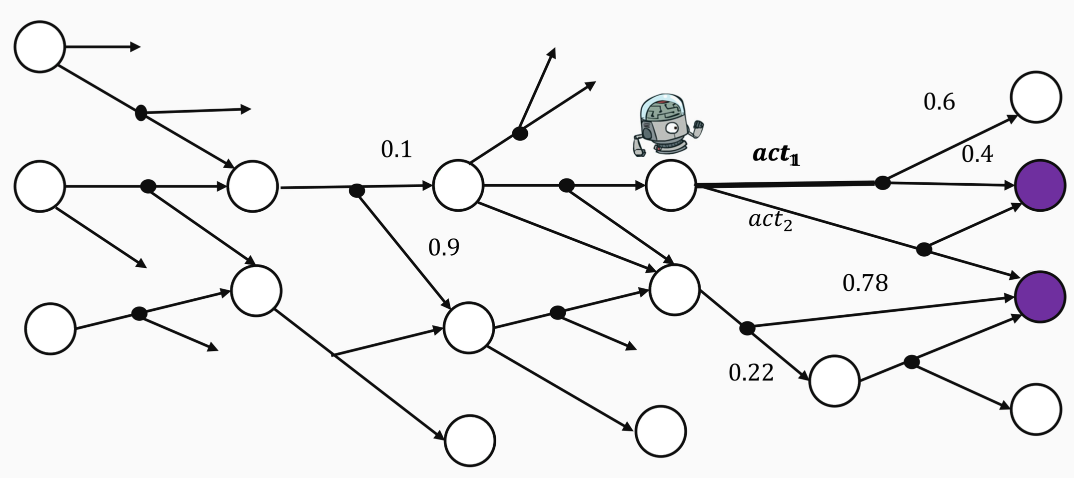
The first question we answer via probabilistic model checking, and the second one we answer by generating counterfactual scenarios.

Model Setting

- **Environment**: *Markov Decision Processes (MDP)* $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P})$.
- **Agent**: Represented as a (deterministic memoryless) policy $\pi : \mathcal{S} \to \mathcal{A}$.
- **Intention**: Set of states $\mathcal{S}_\mathcal{I} \subset \mathcal{S}$ that the agent intends to reach.

Probabilistic Model Checking: What is the probability to reach $\mathcal{S}_\mathcal{I}$?

- $\mathcal{P}_\pi(\texttt{Reach}(\mathcal{S}_\mathcal{I}), s)$: probability to reach $\mathcal{S}_\mathcal{I}$ from $s \in \mathcal{S}$ following policy π.
- $\mathcal{P}_{\max/\min|\Pi}(\texttt{Reach}(\mathcal{S}_\mathcal{I}), s)$: Max./min. probability, for any policy in $\pi \in \Pi$.



Agency and Intention

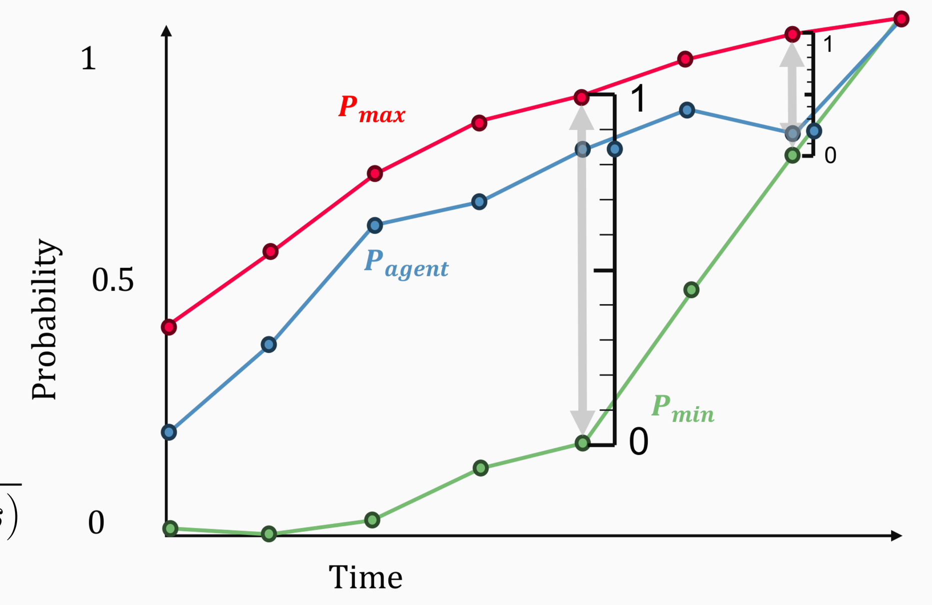
Given an agent π at a state $s \in \mathcal{S}$, we define:

- **Scope of Agency** $(\sigma(s))$: Measures the effect of agent's actions on reaching $\mathcal{S}_\mathcal{I}$.

$$\sigma(s) = \mathcal{P}_{\max|\Pi}(\texttt{Reach}(\mathcal{S}_\mathcal{I}), s) - \mathcal{P}_{\min|\Pi}(\texttt{Reach}(\mathcal{S}_\mathcal{I}), s)$$

- **Intention-quotient** $(\rho_\pi(s))$: Measures how close π is to being optimal to reach $\mathcal{S}_\mathcal{I}$.

$$\rho_\pi(s) = \frac{\mathcal{P}_\pi(\texttt{Reach}(\mathcal{S}_\mathcal{I}), s) - \mathcal{P}_{\min|\Pi}(\texttt{Reach}(\mathcal{S}_\mathcal{I}), s)}{\mathcal{P}_{\max|\Pi}(\texttt{Reach}(\mathcal{S}_\mathcal{I}), s) - \mathcal{P}_{\min|\Pi}(\texttt{Reach}(\mathcal{S}_\mathcal{I}), s)}$$



For a sequence of states (or trace) $\tau = (s_1, \dots, s_n)$:

Scope of agency $(\overline{\sigma}(\tau))$: Average along a sequence of events (states) τ of the scope of agency.

$$\overline{\sigma}(\tau) = \frac{1}{|\tau|} \sum_{s \in \tau} \sigma(s).$$

Intention-quotient $(\overline{\rho}_\pi(\tau))$: Weighted average along trace τ, weighted by the scope of agency.

$$\rho_\pi(\tau) = \frac{1}{\sum_{s\in\tau} \sigma(s)} \sum_{s\in\tau} \sigma(s)\rho_\pi(s)$$

Methodology: Analysis of Intentional Behavior, using Counterfactual Reasoning to Augment Evidence

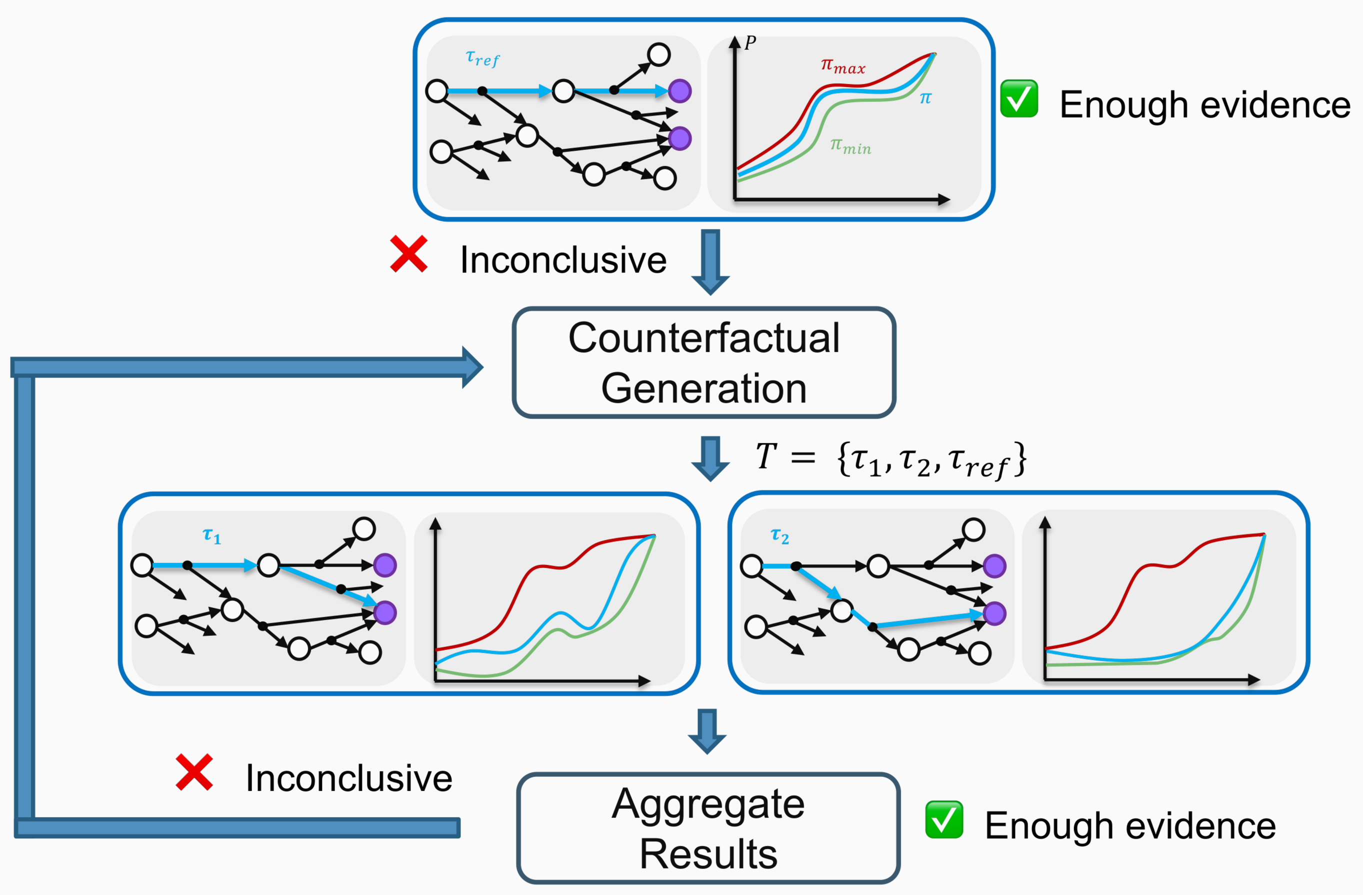
Setting: A factual trace τ that reaches $\mathcal{S}_\mathcal{I}$ has happened. We want to analyze whether the agent π reached $\mathcal{S}_\mathcal{I}$ intentionally or not. Because of uncertainty, we can only determine *evidence* of intentional behavior. Guiding principle:

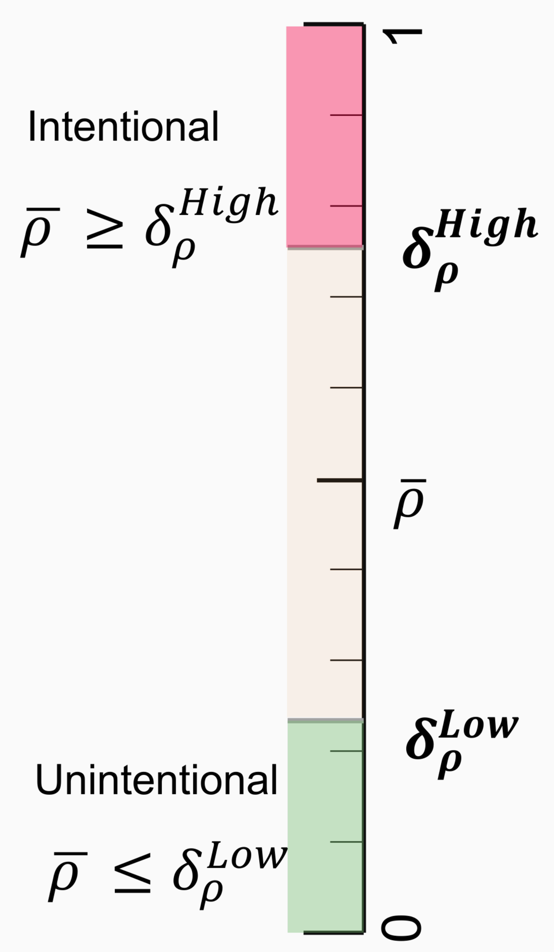
- The agent behaves closely to maximizing probability to reach $\mathcal{S}_\mathcal{I}$,
- The agent could have behaved otherwise.

Thresholds on evidence

- *Agency threshold* $(\boldsymbol{\delta_\sigma})$. Scope of agency along a trace needs to be larger than the threshold, i.e., $\overline{\sigma}(\tau) \geq \delta_\sigma$. Otherwise, more evidence is required.
- *Intention thresholds* $(\boldsymbol{\delta_\rho^{High}}, \boldsymbol{\delta_\rho^{Low}})$.
  - If $\overline{\rho}_\pi(\tau) \geq \delta_\rho^{High}$, the reaching of $\mathcal{S}_\mathcal{I}$ is considered *intentional*.
  - If $\overline{\rho}_\pi(\tau) \leq \delta_\rho^{Low}$, the reaching of $\mathcal{S}_\mathcal{I}$ is considered *unintentional*.
  - Otherwise, more evidence is required.

Counterfactual traces: New counterfactual traces are generated if there is not enough evidence to establish intentionality. Agency and intention-quotient are aggregated along new traces, and compared against thresholds.





Counterfactual Generation

Relevant counterfactuals should be generated with a human in the loop. We propose two semi-automatic generation techniques:

- **Factored MDP**. State space factored into *integral* and *peripheral* state variables $S = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$. Generate counterfactuals sampling integral variables.
- **Distances on MDPs**. Define a distance notion on states of the MDP, sample traces at close distance.

Discussion & Future Work

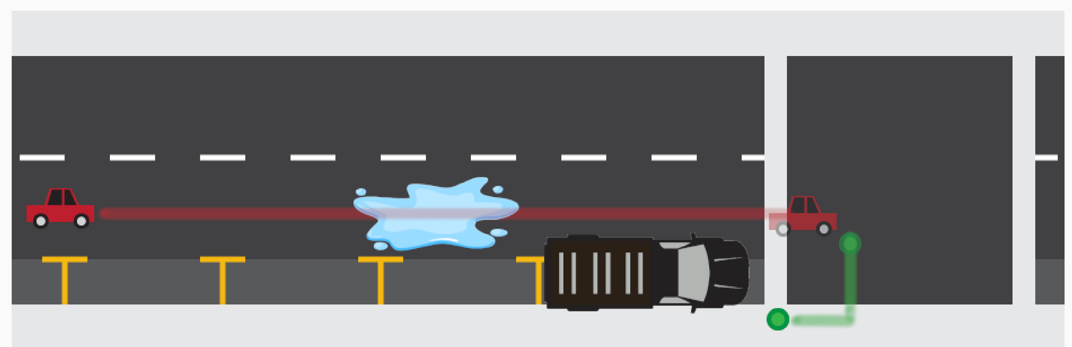
- **Limitations**:
  - Need for an MDP model of the environment and the agent.
  - Probabilistic model checking is costly.
  - Agent's beliefs are not taken into account.
- **Future directions**:
  - *General policies.* Extending to policies with memory and non-determinism is feasible, although computationally more expensive.
  - *Multi-agent setting.* Considering several agents that interact towards shared or conflicting goals.
  - *Time extension.* Longer traces, study intention reconsideration.

Case Study

Example:

- An autonomous car collided with a pedestrian.
- A section of the road was slippery, and there was a truck blocking visibility.

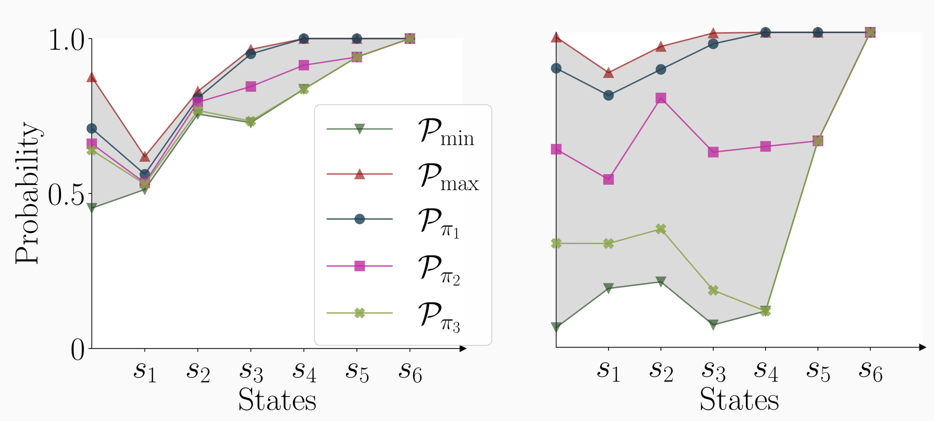
Was the collision intentional?



Charateristics of the MDP:

- *States:* Position of pedestrian with respect to the car, speed of the car, discretized to integers of m and ms^{-1}.
- *Actions:* Accelerate, brake, coast.
- *Size:* 120k states, 400k transitions.

Generation of counterfactuals: Sample variations of pedestrian behavior, visibility, slippery range, and friction.



Comparative analysis of several agents: We built three agents and studied them on the same trace.

- Agent π_1 (—●—) drives to intentionally hit the pedestrian.
- Agent π_2 (—■—) drives as fast as possible, caring very little for the safety of the pedestrian.
- Agent π_3 (—✖—) drives in a safe manner.