Trust Topologies in Verification of Social Explainable AI

Damian Kurpiewski

Institute of Computer Science
Polish Academy of Sciences

Faculty of Mathematics and Computer Science
Nicolaus Copernicus University in Toruń

(joint work with Wojciech Jamroga, Łukasz Mikulski and Teofil Sidoruk)

BLESS, 23/11/2023

ATL: What Agents Can Achieve

- ATL: Alternating-time Temporal Logic [Alur et al. 1997-2002]
- Temporal logic meets game theory
- Main idea: cooperation modalities

$\langle\langle A \rangle\rangle \Phi$: coalition A has a collective strategy to enforce Φ

$\rightsquigarrow$ Φ can include temporal operators: X (next), F (sometime in the future), G (always in the future), U (strong until)

Semantic Variants of ATL

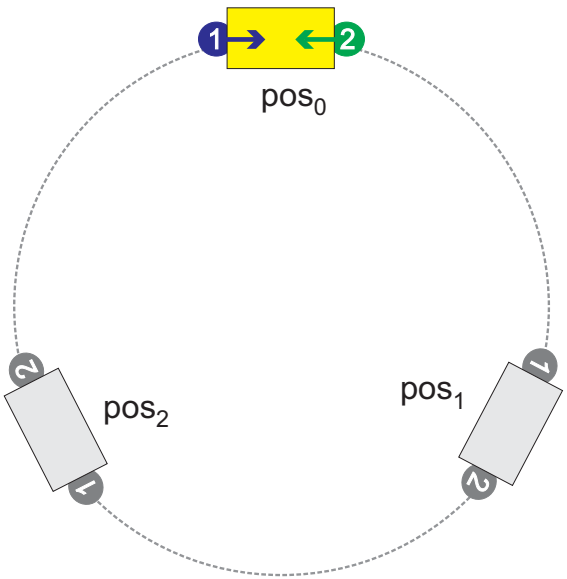
Memory of agents:

- Perfect recall (R) vs. imperfect recall strategies (r)

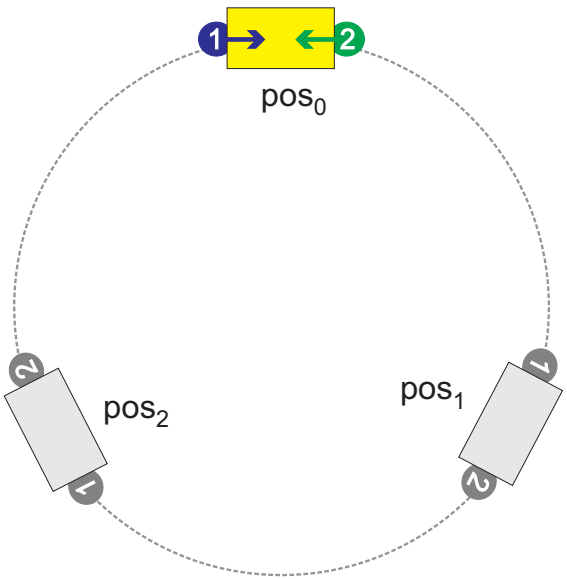
Available information:

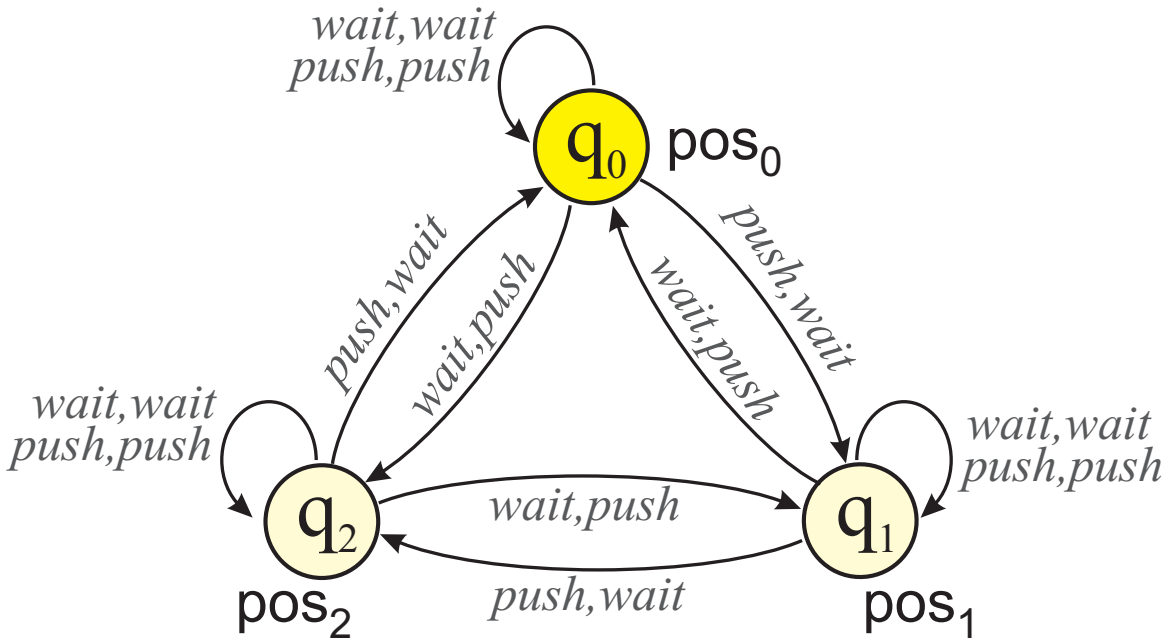
- Perfect information (I) vs. imperfect information strategies (i)

Example: Robots and Carriage

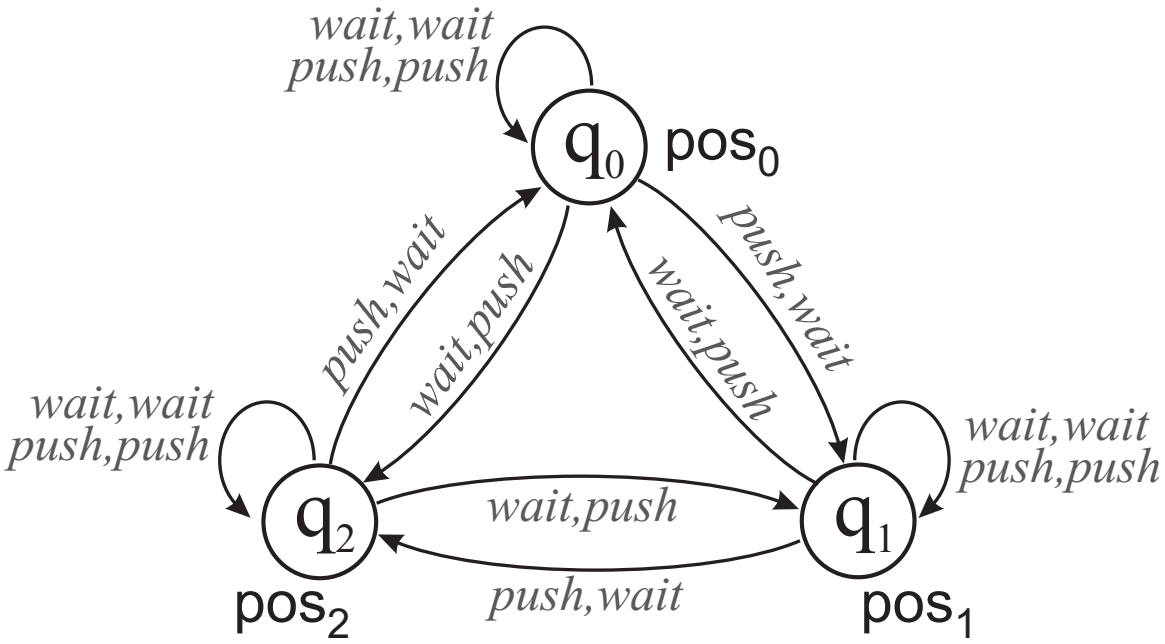


Example: Robots and Carriage



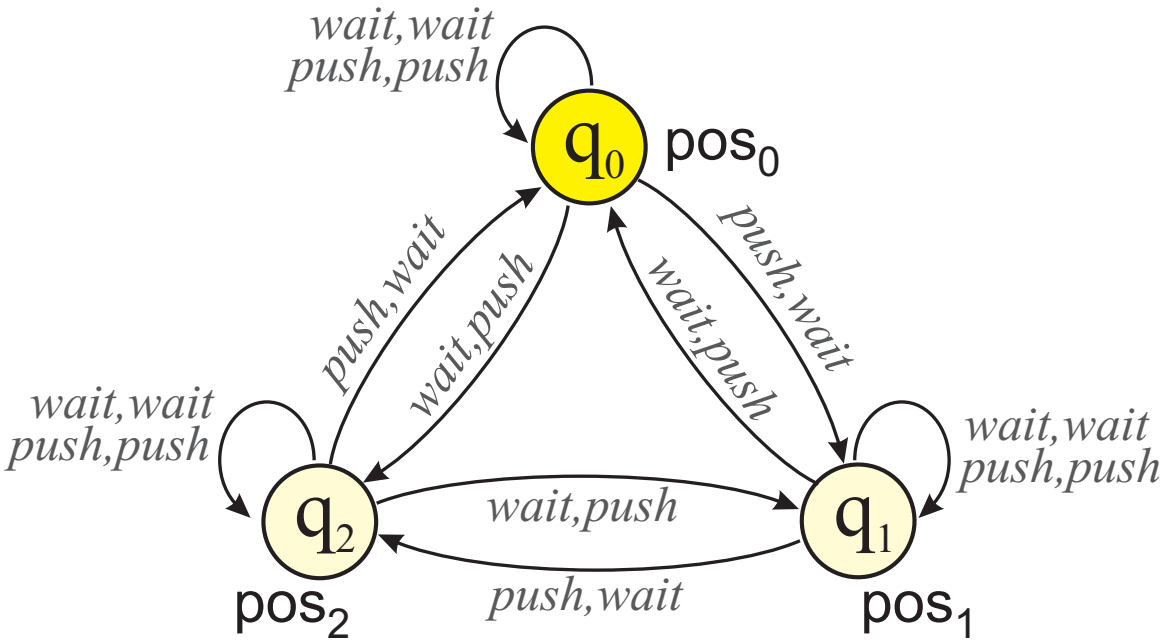


Example: Robots and Carriage



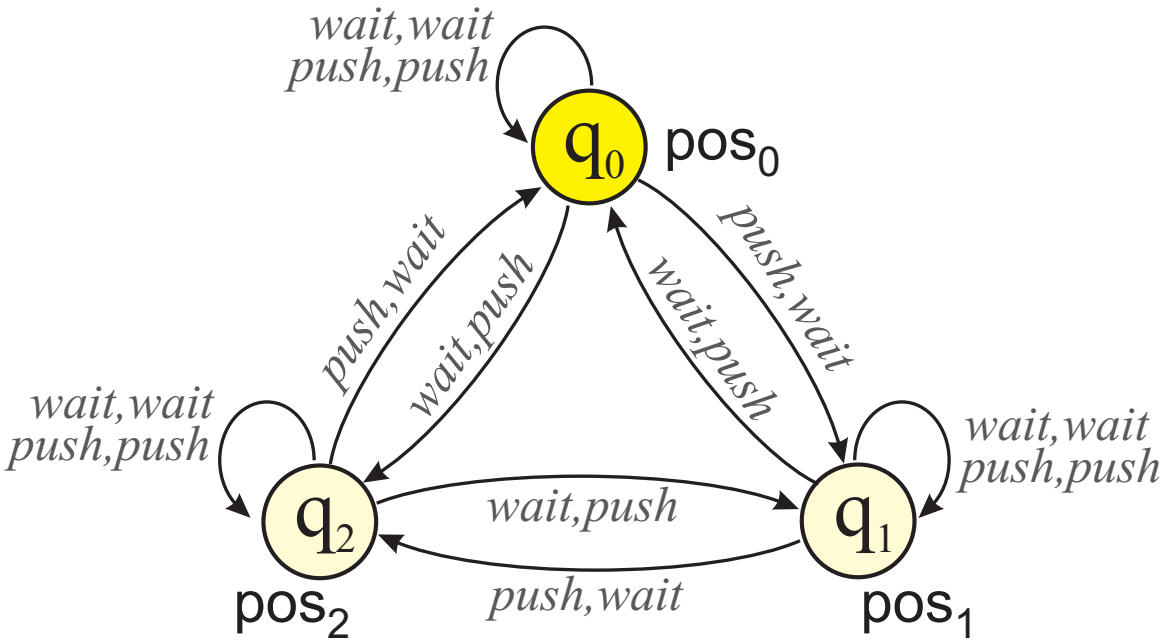
$\text{pos}_0 \rightarrow \langle\langle 1 \rangle\rangle \text{F} \text{pos}_1$

Example: Robots and Carriage



$pos_0 \rightarrow \langle\langle 1 \rangle\rangle F pos_1$

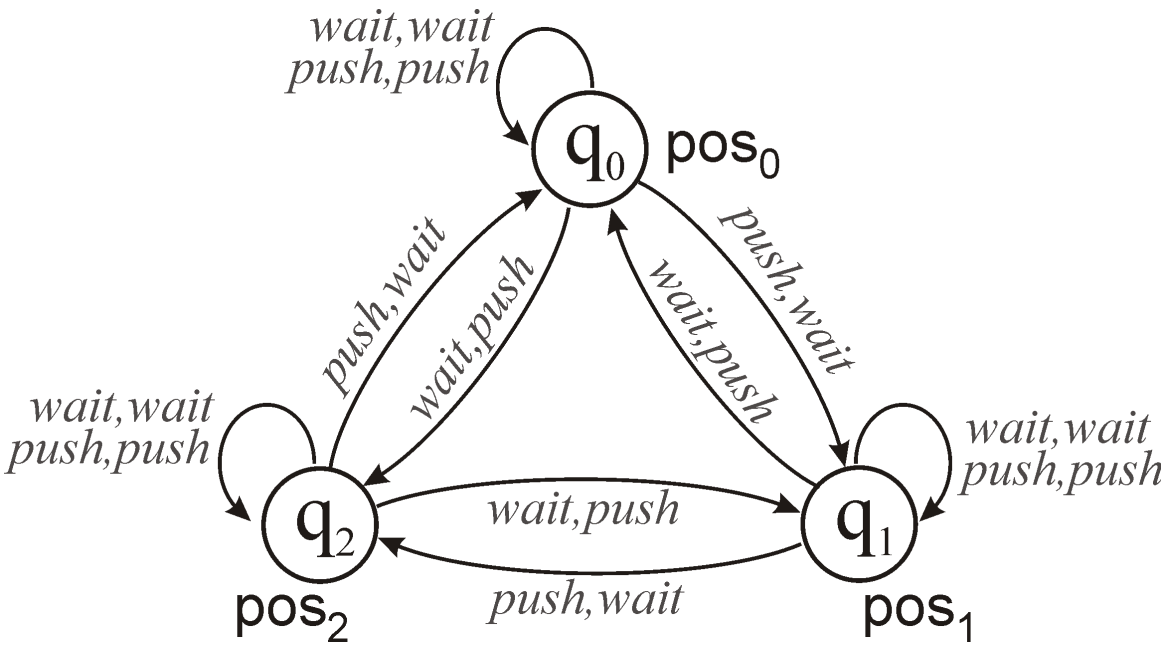
Example: Robots and Carriage



$pos_0 \rightarrow \langle\langle 1 \rangle\rangle F pos_1$

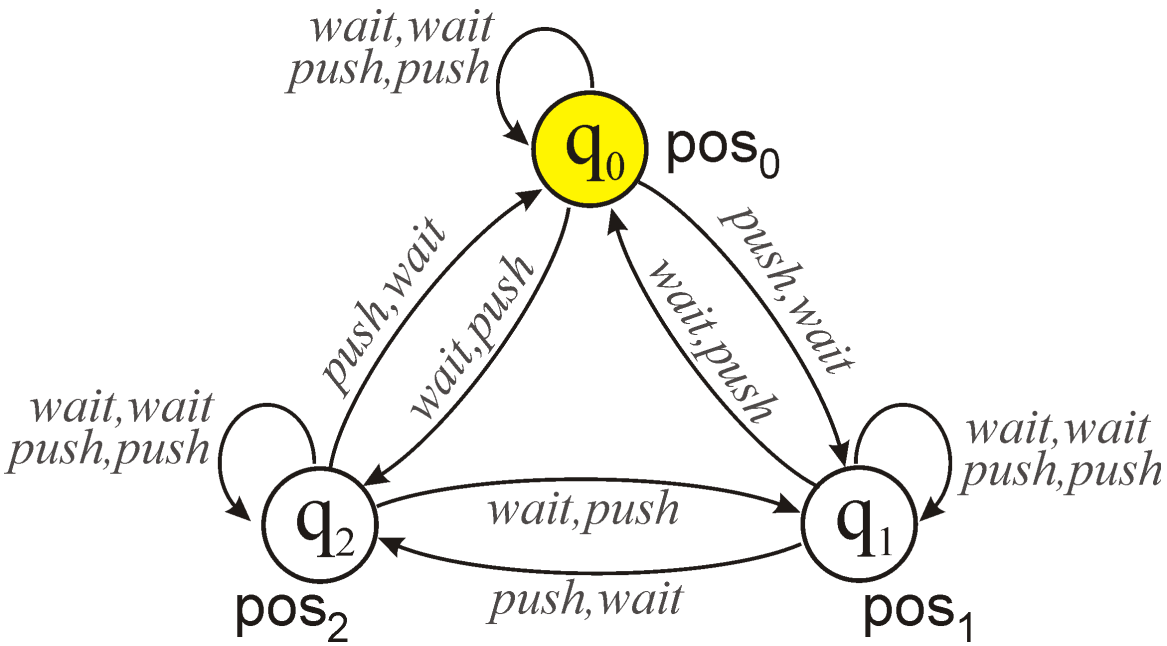
No!

Example: Robots and Carriage



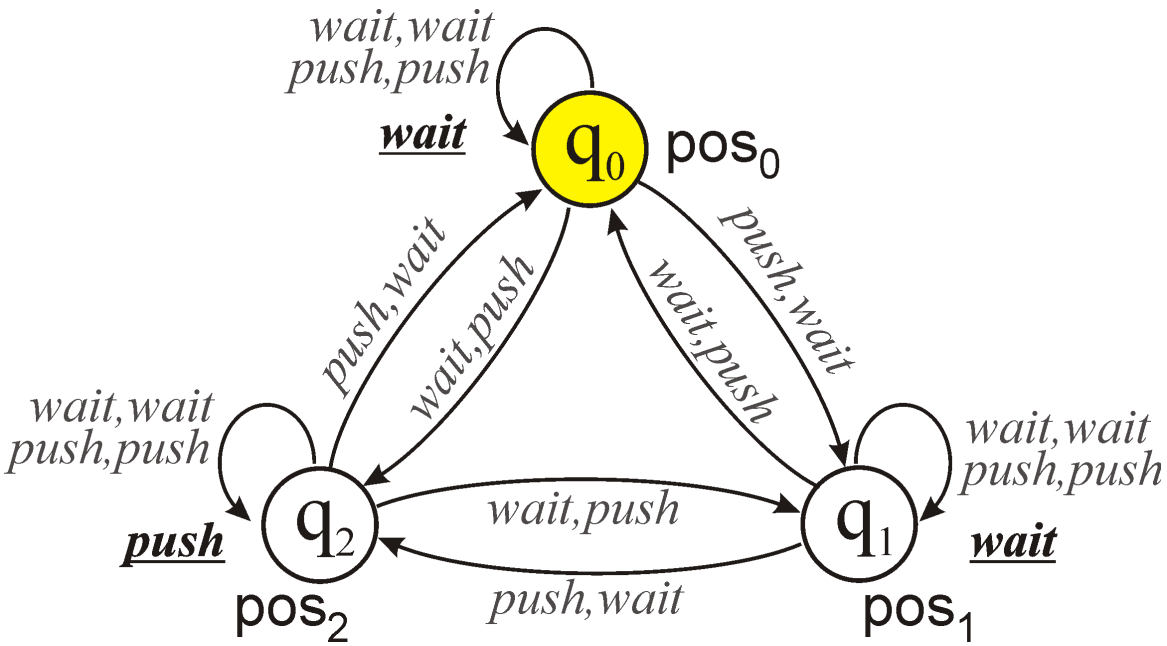
$pos_0 \rightarrow \langle\langle 1 \rangle\rangle G \neg pos_1$

Example: Robots and Carriage



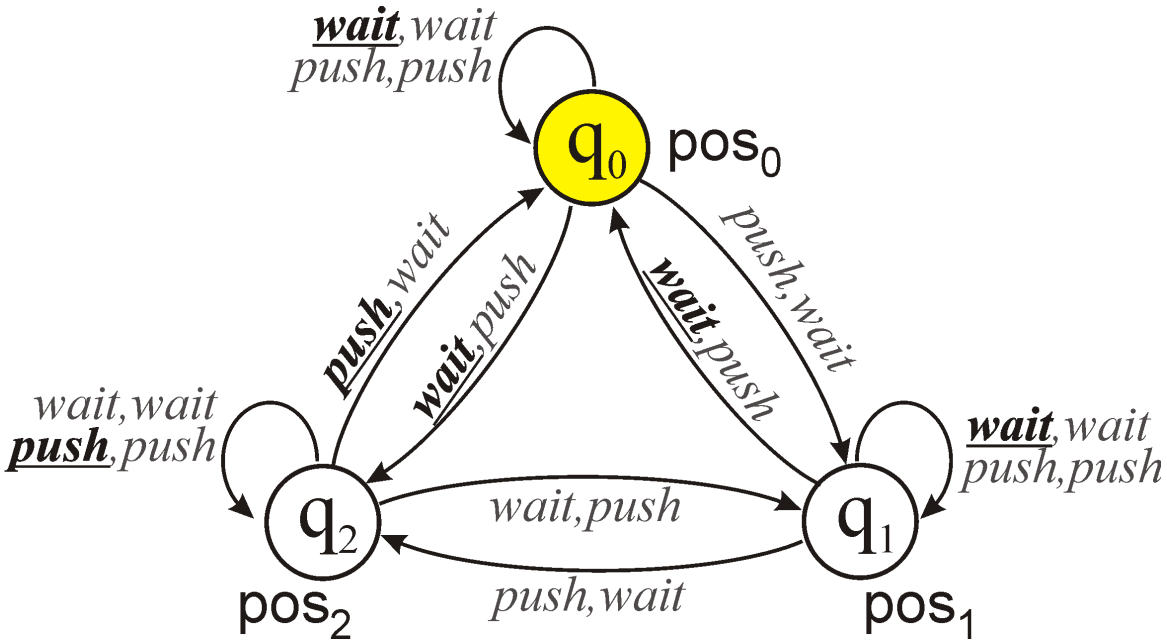
$\mathrm{pos}_0 \rightarrow \langle\langle 1 \rangle\rangle \mathrm{G} \neg \mathrm{pos}_1$

Example: Robots and Carriage



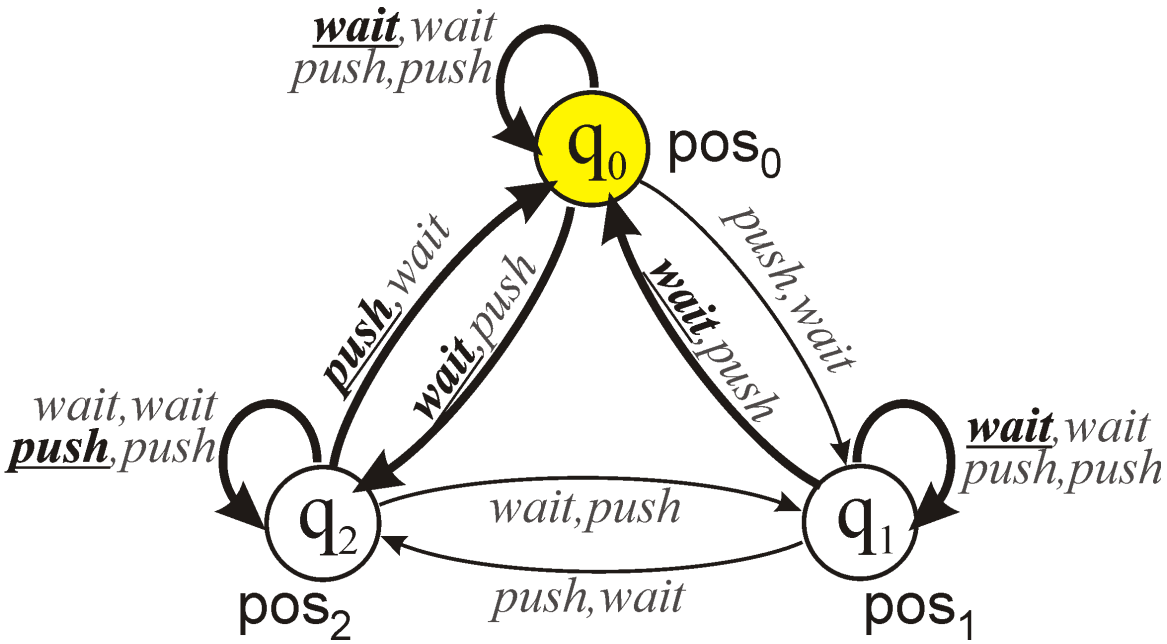
$pos_0 \rightarrow \langle\langle 1 \rangle\rangle G \neg pos_1$

Example: Robots and Carriage



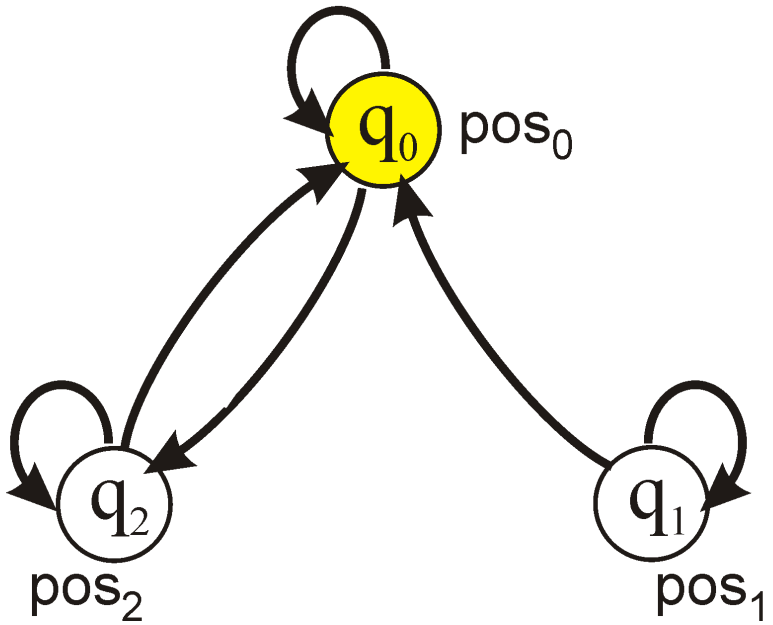
$pos_0 \rightarrow \langle\langle 1 \rangle\rangle G \neg pos_1$

Example: Robots and Carriage



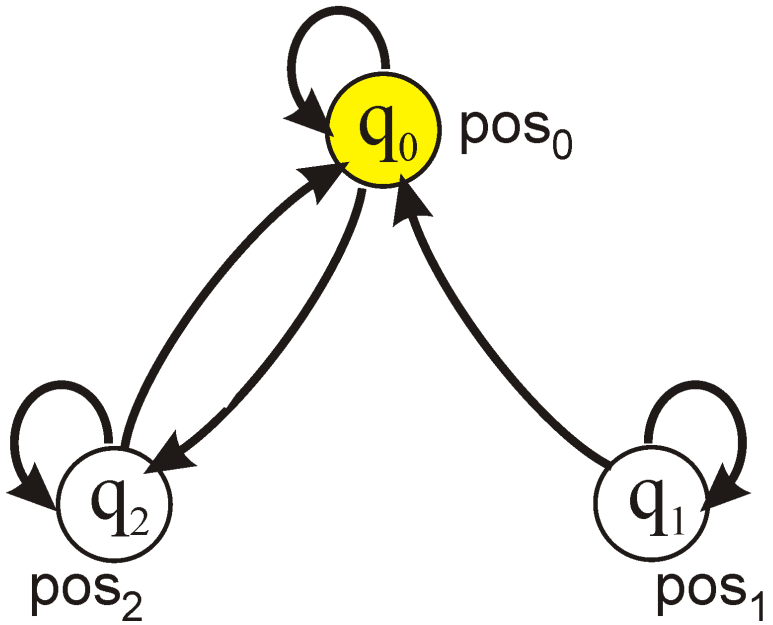
$\mathrm{pos}_0 \rightarrow \langle\langle 1 \rangle\rangle \mathrm{G} \neg \mathrm{pos}_1$

Example: Robots and Carriage



$pos_0 \rightarrow \langle\langle 1 \rangle\rangle \mathrm{G} \neg pos_1$

Example: Robots and Carriage



$pos_0 \rightarrow \langle\langle 1 \rangle\rangle \mathrm{G} \neg pos_1$

Yes!

ATL with incomplete information

- Imperfect information ($q \sim_a q'$)

ATL with incomplete information

- Imperfect information ($q \sim_a q'$)
- Imperfect recall - agent memory coded within state of the model

ATL with incomplete information

- Imperfect information ($q \sim_a q'$)
- Imperfect recall - agent memory coded within state of the model
- Uniform strategies - specify same choices for indistinguishable states: $q \sim_a q' \implies s_a(q) = s_a(q')$

ATL with incomplete information

- Imperfect information ($q \sim_a q'$)
- Imperfect recall - agent memory coded within state of the model
- Uniform strategies - specify same choices for indistinguishable states: $q \sim_a q' \implies s_a(q) = s_a(q')$
- Fixpoint equivalences **do not hold** anymore

ATL with incomplete information

- Imperfect information ($q \sim_a q'$)
- Imperfect recall - agent memory coded within state of the model
- Uniform strategies - specify same choices for indistinguishable states: $q \sim_a q' \implies s_a(q) = s_a(q')$
- Fixpoint equivalences **do not hold** anymore
- Model checking $\mathbf{ATL}_{\mathrm{ir}}$ is Δ_2^p-complete

SAI

- A novel approach in AI focusing on decentralization and transparency.
- Emphasizes the social context and human-centricity in AI applications.
- Aims to overcome the black-box nature of traditional AI systems.
- Moving from centralized control to **individualized AI entities** that interact with each other.
- Incorporating explainability by design, fostering trust and understanding in AI systems.

Formal Modelling of SAI

Multi-Agent Systems (MAS)

Networks of agents that interact with each other to achieve certain goals.

Asynchronous Multi-Agent Systems (AMAS)

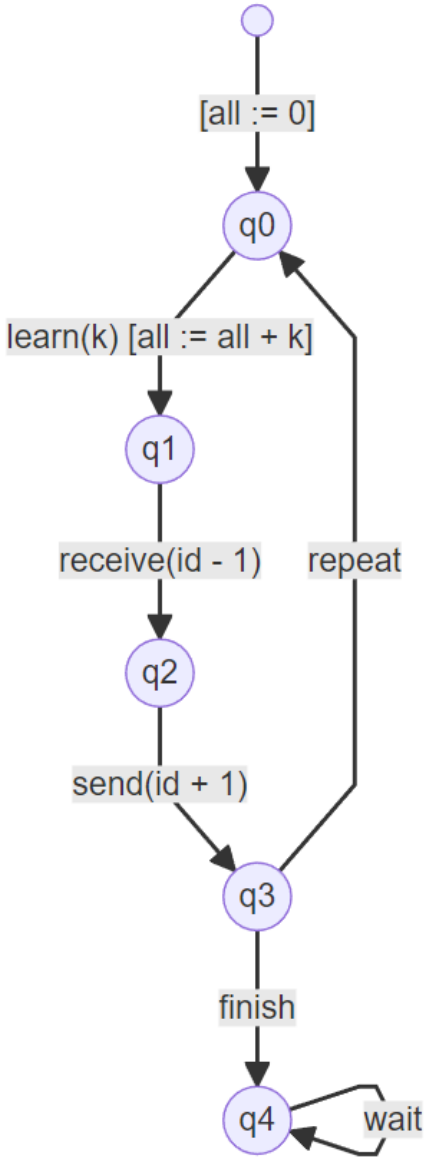
A type of MAS where agents operate and interact asynchronously, allowing for more complex and realistic modeling of systems.

- Utilizing AMAS to **model the network of Personal AI Valets** (PAIVs) and **formalize their properties** using logical frameworks.
- Modeling the network of PAIVs as an AMAS to capture the dynamics of decentralized, interactive, and explainable AI environments.

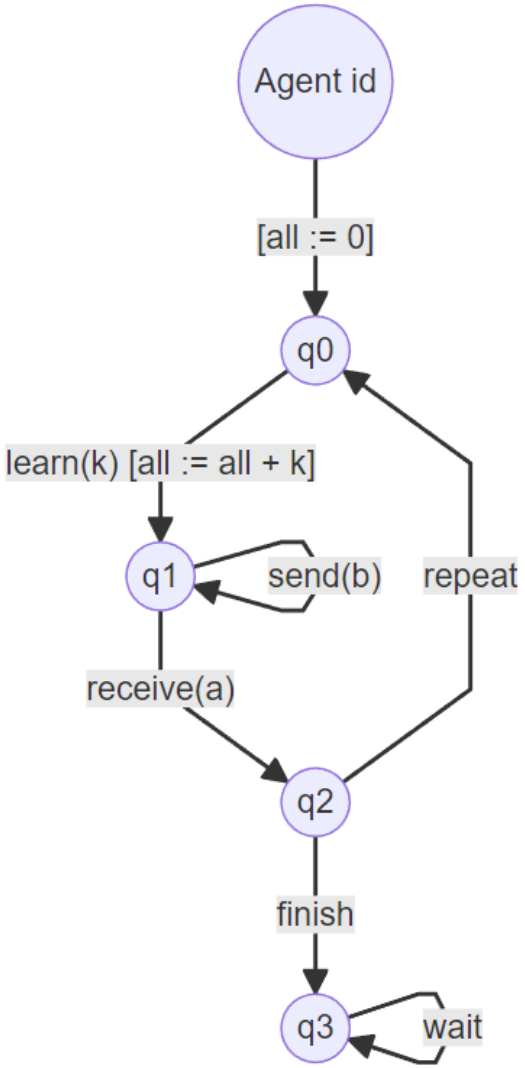
Modelling Agents

- Each agent represents a PAIV
- Focus on sharing phase to facilitate interaction and collaboration among agents.
- Sharing: the phase where the agent shares its findings and collaborates with other agents in the network.
- Order of interactions between agents based on the underlying network-like topology.
- Using **trust topologies** to define trust relationships between agents.

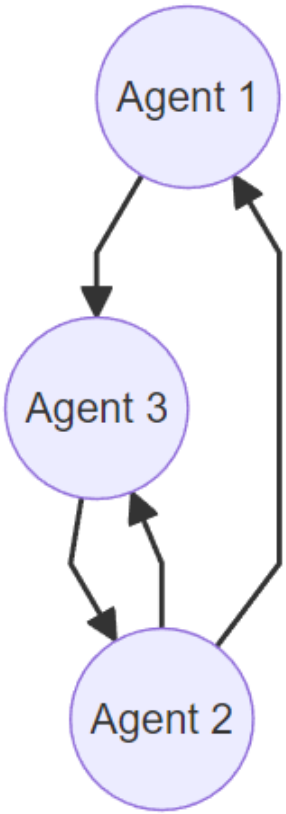
Example: Ring Topology



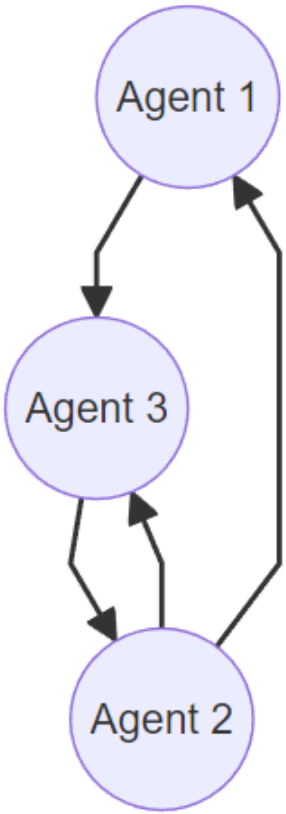
Example: Universal Template

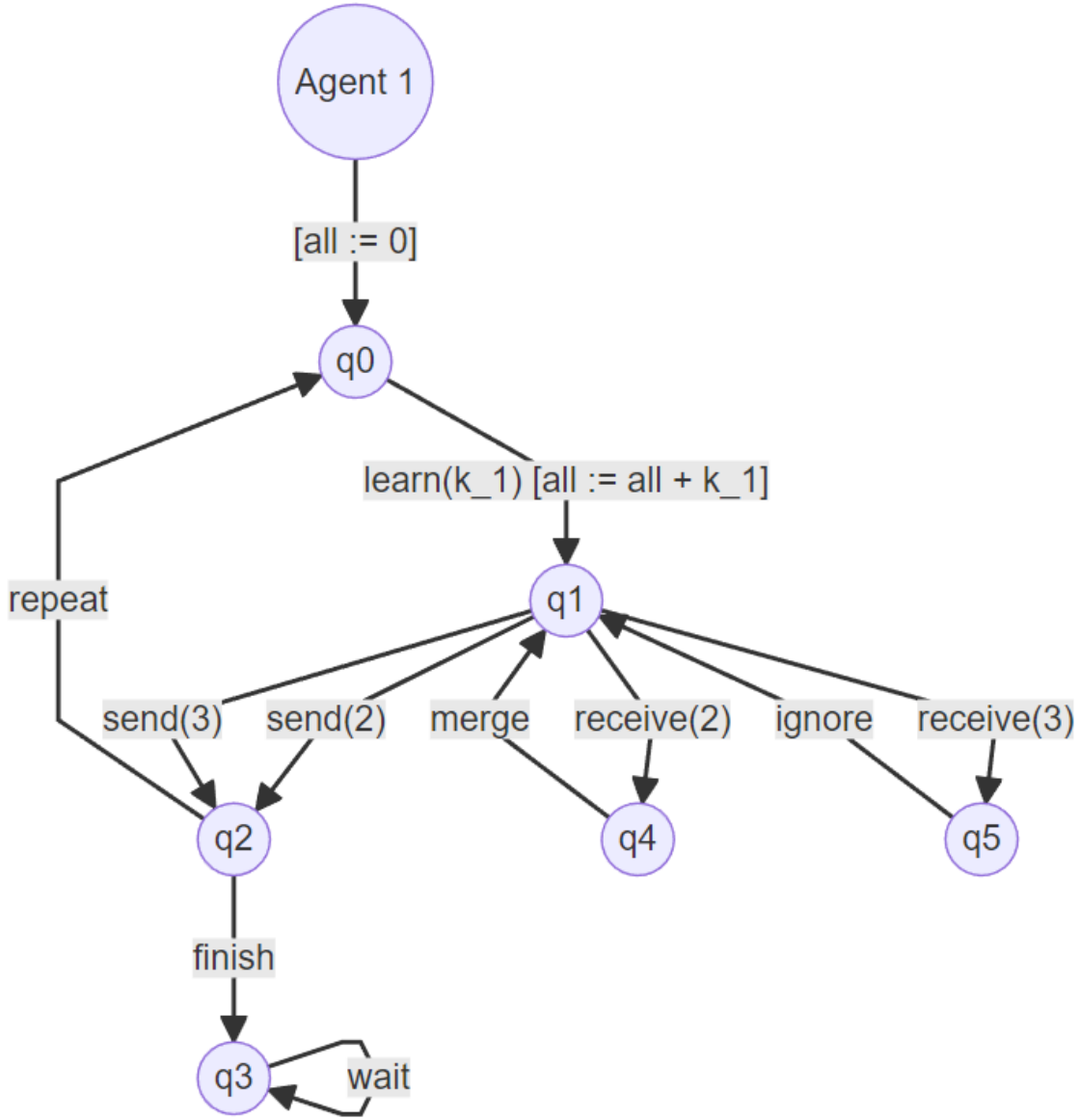


Example: Trust Topology (3 Agents)



Example: Trust Topology (3 Agents)





Attack Scenarios

Man in the Middle

- Engages actively during the sharing phase.
- Has the ability to intercept any model being transmitted by one of the honest agents and subsequently relay it to another agent.

Impersonator

- Involves an AI agent being compromised with malicious code, leading to undesirable behavior.
- Adheres to the sharing protocol when disseminating its model to others.
- Possesses the capability to falsify the quality of its local AI model, thereby deceiving the subsequent agent into accepting it.

Our Goals

- Specify and implement models of Social Explainable AI (SAI) protocol, with a focus on the sharing phase.
- Define essential properties of these models, specify them in Alternating-time Temporal Logic with Imperfect Recall (**ATL**$_{\text{ir}}$), and verify them using our specialized model-checker.
- Explore how different trust topologies and sharing protocols can enhance transparency and user trust in SAI systems.
- Develop simulation environments to evaluate the effectiveness of the SAI protocol in various contexts.

Damian Kurpiewski, Wojciech Jamroga, Teofil Sidoruk:
Towards Modelling and Verification of Social Explainable AI.
ICAART 2023: 396-403

Thank You