Reward Design for Justifiable Sequential Decision-Making

Aleksa Sukovic[1,2], Goran Radanovic[1]

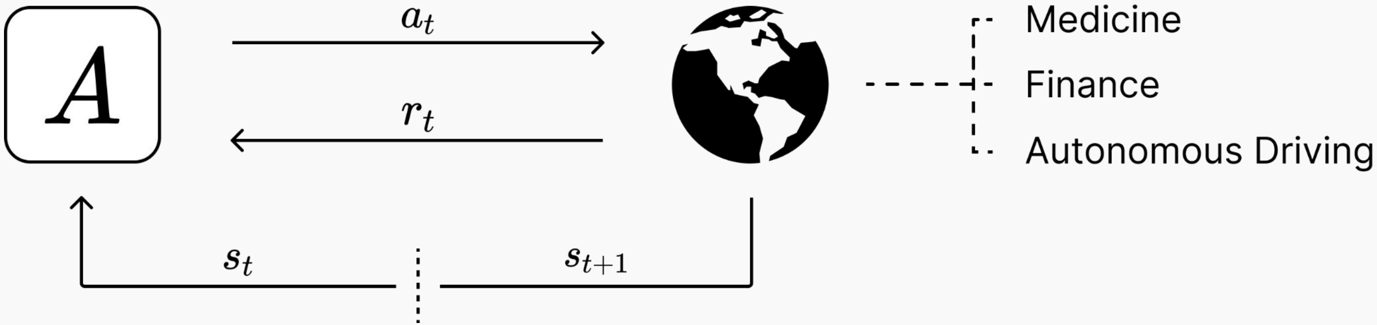
Max Planck Institute for Software Systems, Saarland University

{asukovic,gradanovic}@mpi-sws.org

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

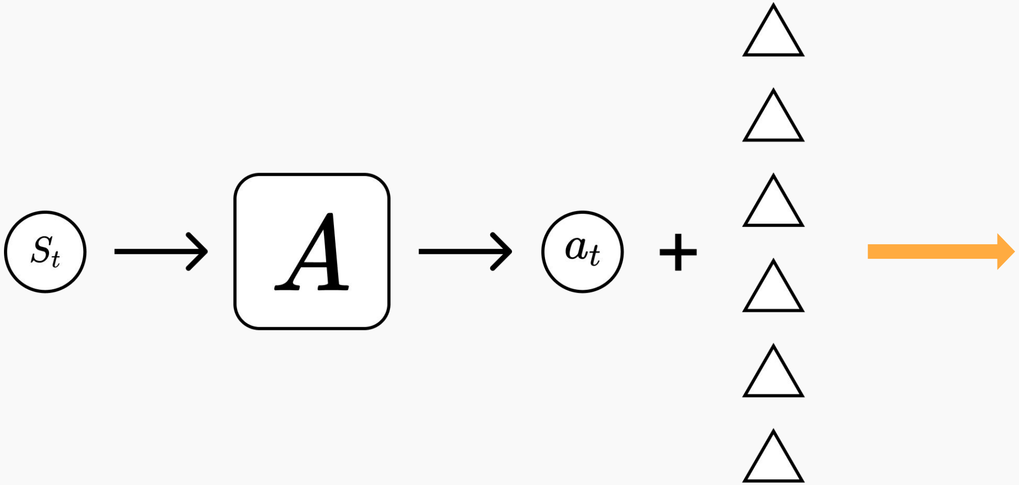
UNIVERSITÄT
DES
SAARLANDES

Justifiability Is Necessary



- Medicine
- Finance
- Autonomous Driving

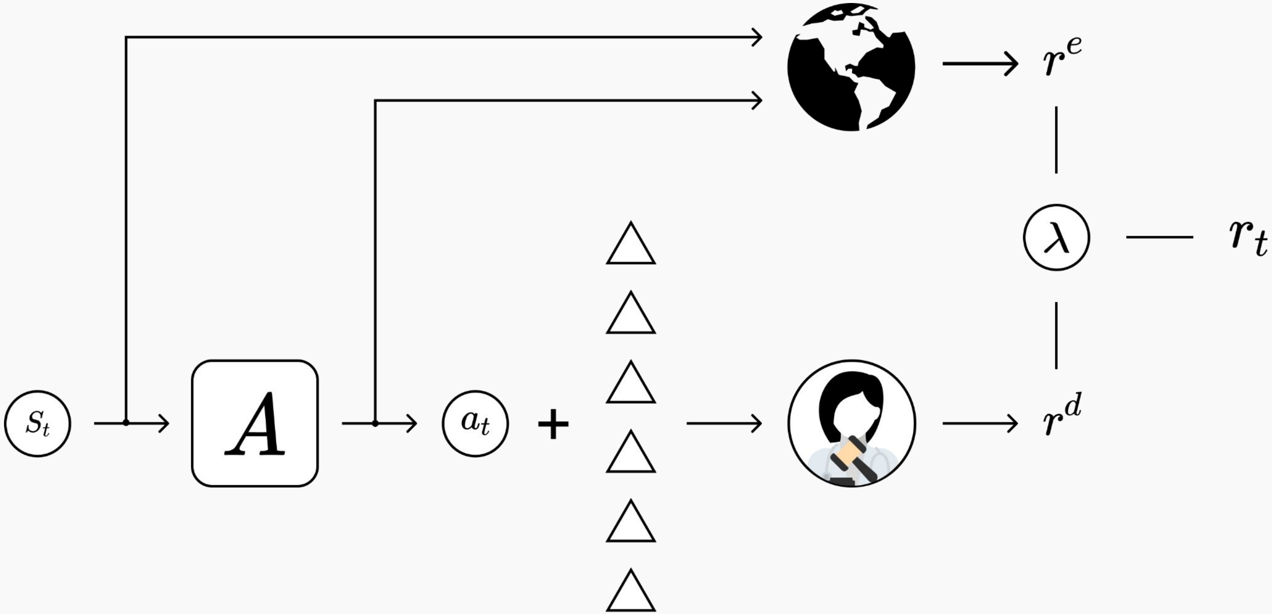
Justifiability Is Necessary



a decision needs to be **justified** with **supporting evidence**

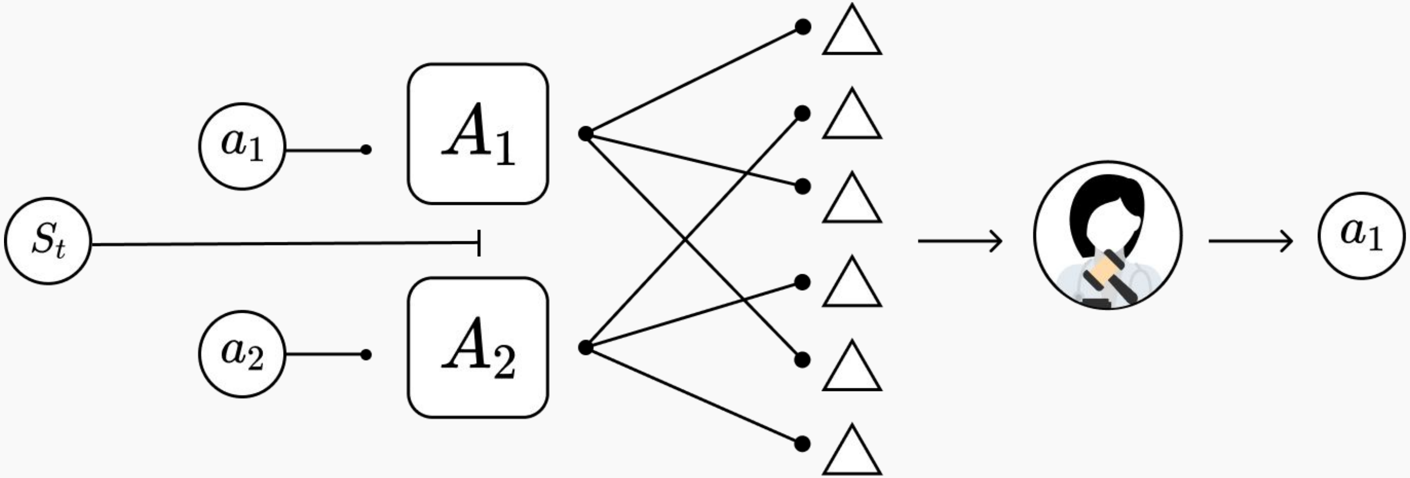
This Work

How can we design **rewards** that incentivize the agent to **complete** a **task**, but **through decisions** that can be **justified** with supporting evidence?



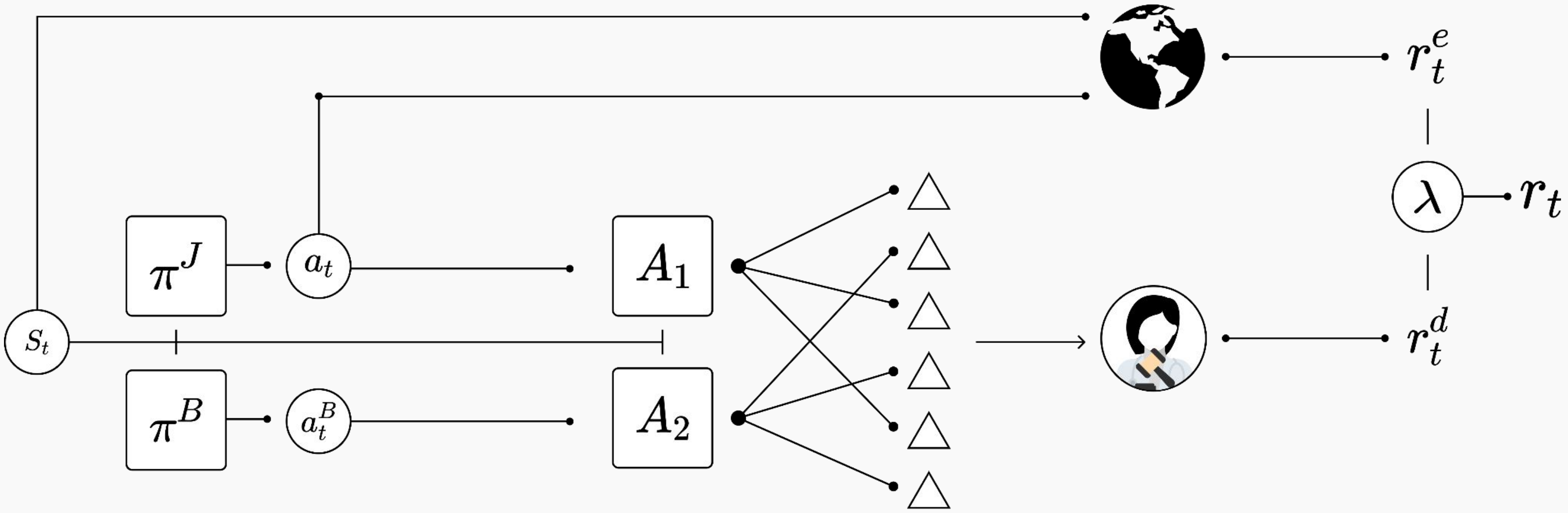
Debate as an Interpretable Justifiability Reward

Use the outcome of a zero-sum **debate game** as a **justifiability reward**



Debate as an Interpretable Justifiability Reward

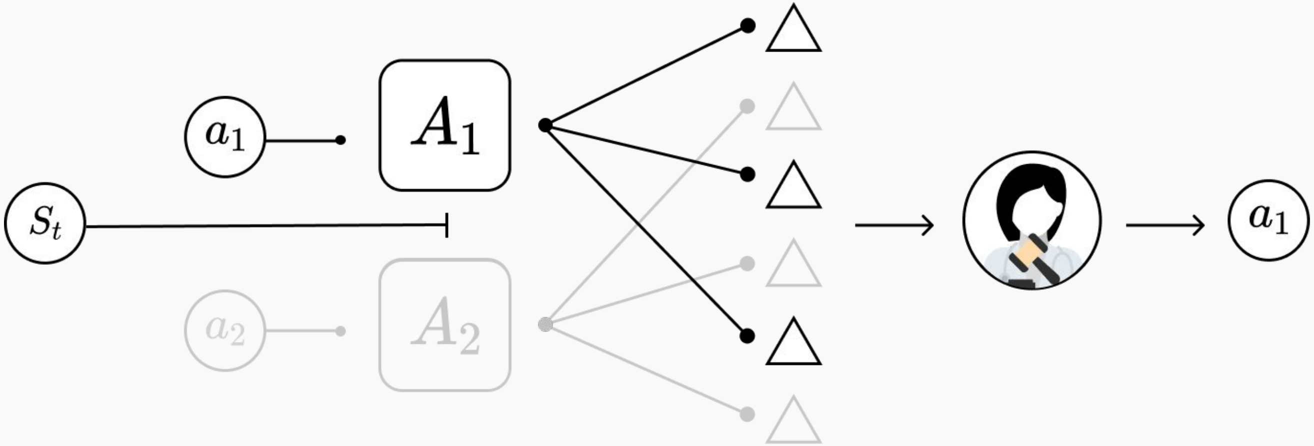
Aim to improve a **baseline** policy that only optimizes for environment rewards



Learning to Propose Evidence

Treat debate as an instance of a **contextualized** extensive-form game

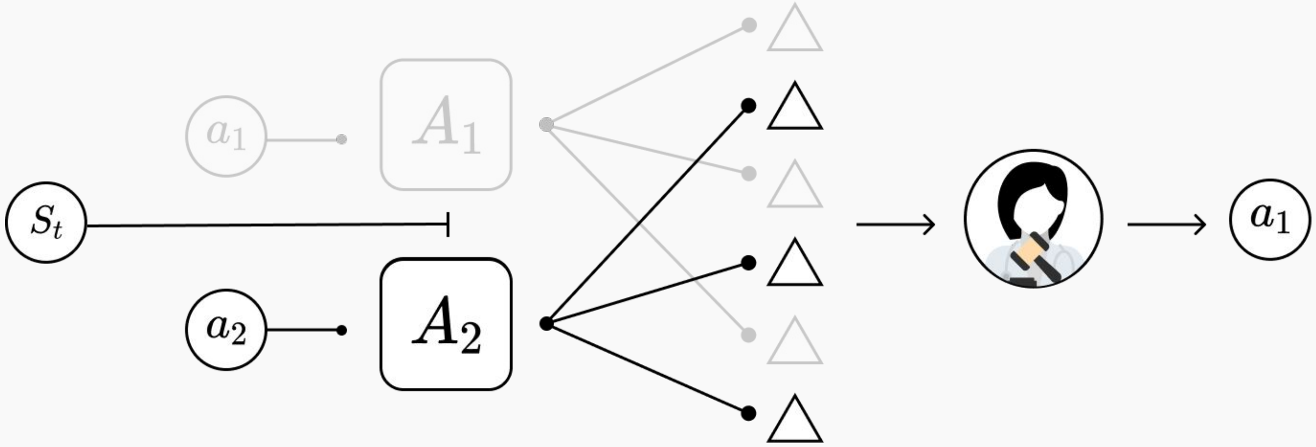
Maxmin approach



Learning to Propose Evidence

Treat debate as an instance of a **contextualized** extensive-form game

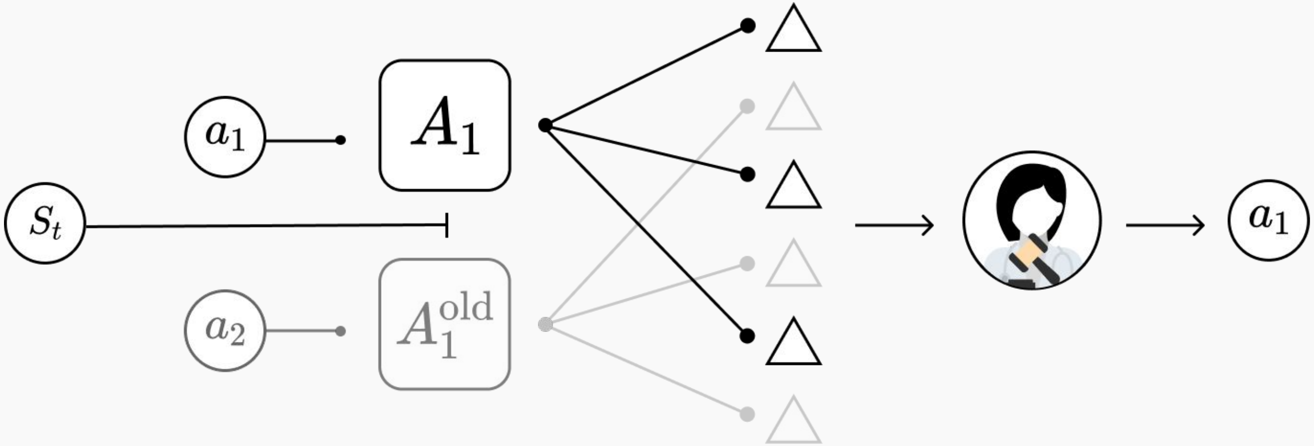
Maxmin approach



Learning to Propose Evidence

Treat debate as an instance of a **contextualized** extensive form game

Self-play approach



Justifying Decisions in a Healthcare Setting

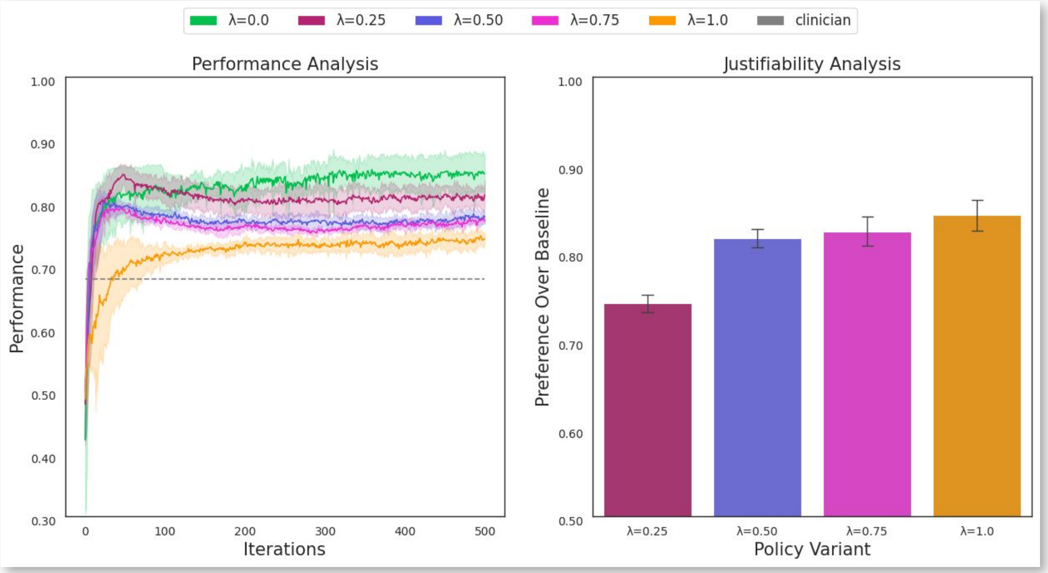
- **MIMIC-III** dataset, extract ~*18,000* unique patients
- **State-** and **evidence-space** is *continuous* and *44-dimensional*
- We set **number** of **turns** to 6 (13.6% of the full state) in all debate games
- 5 choices for **IV** and **VC** medication, **25-dim** discrete **action-space**
- Learn **justifiable policies** using a (deep, double, dueling) **Q-Learning** algorithm

$$\mathcal{D} = \{(s_t, a_{\mathrm{p}}, a_{\mathrm{np}})\}$$

preference points to the
clinician's decision

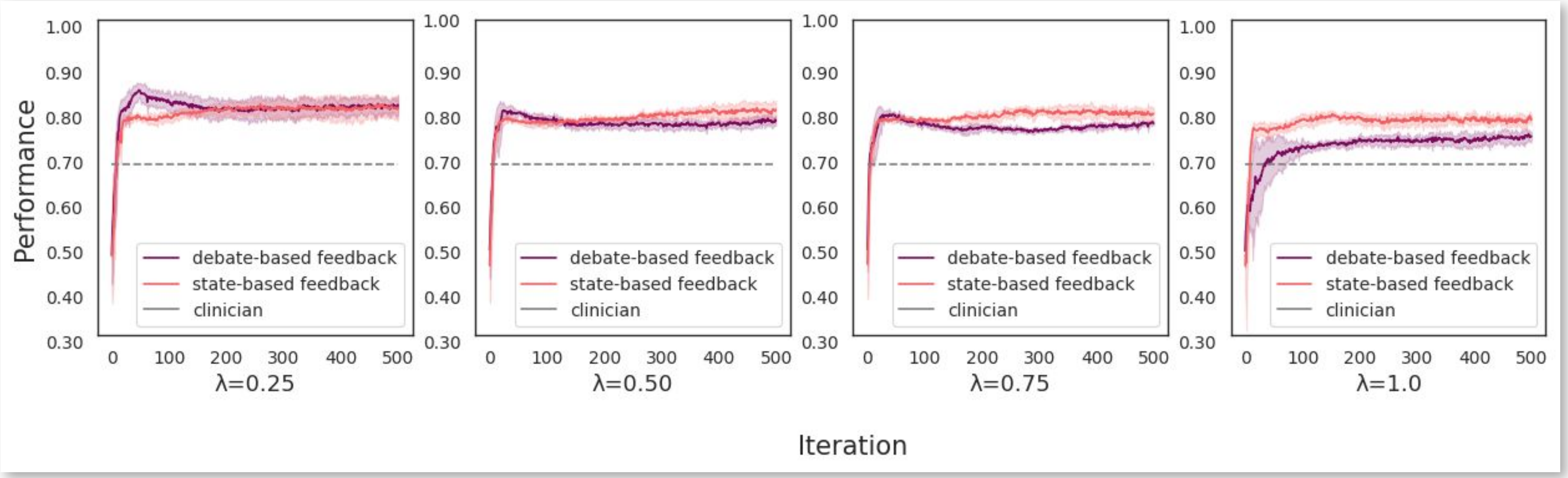
Experiment 1: Effectiveness of Task Policies

Moderate inclusion of the **justifiability reward** yields policies **highly preferred** by the judge



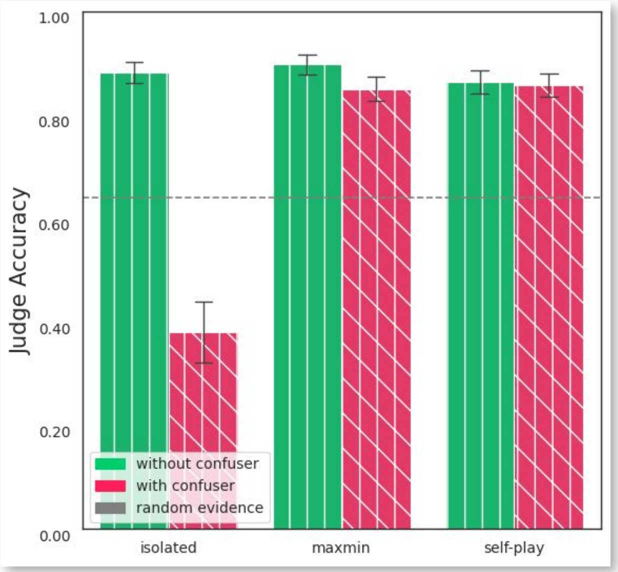
Experiment 2: Debate- vs. State-Based Feedback

Debate enables **good performance** while only exposing the judge to the **13%** of the **state**



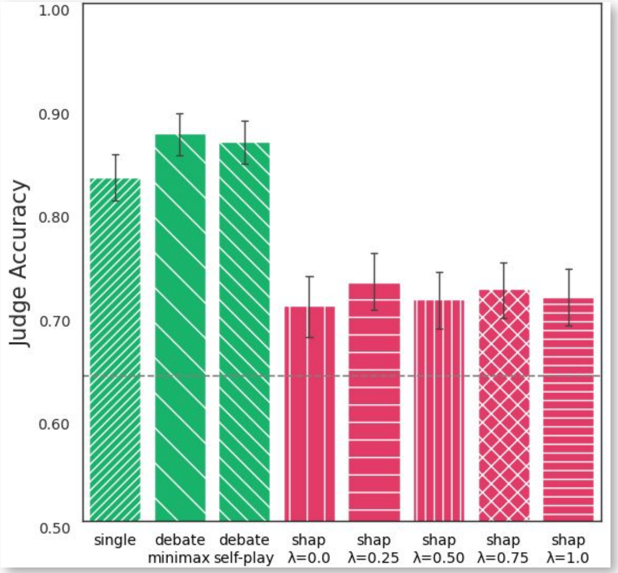
Experiment 3: Effectiveness of Argumentative Policies

Debate agents are both **helpful** and **robust**



Experiment 4: Comparison to SHAP-Based Explanations

SHAP (Shapley additive explanations) are **not as effective** for justifying decisions



Reward Design for Justifiable Sequential Decision-Making

Aleksa Sukovic[1,2], Goran Radanovic[1]

Max Planck Institute for Software Systems, Saarland University

{asukovic,gradanovic}@mpi-sws.org

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

UNIVERSITÄT
DES
SAARLANDES