# Mitigating Negative Side Effects via Environment Shaping

Sandhya Saisubramanian and Shlomo Zilberstein

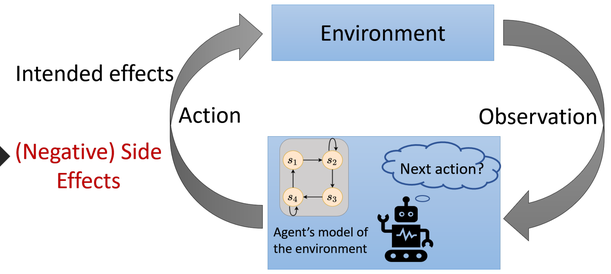
College of Information and Computer Sciences, University of Massachusetts Amherst, USA

## Negative Side Effects

Unanticipated, undesirable consequences of agent actions

Sources of Model Incompleteness

- Unavailability of information during system design
- Unanticipated domain characteristics
- Cultural differences between target users and development team



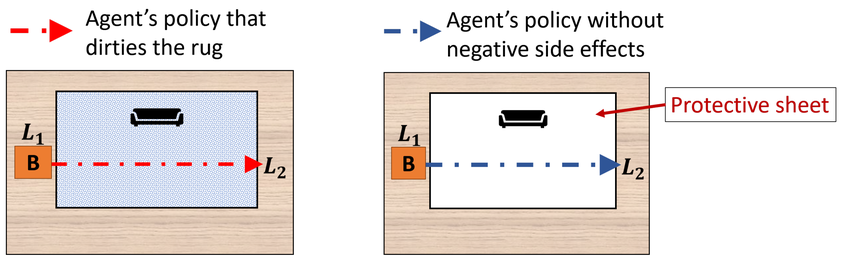

**Goal:** To mitigate negative side effects by leveraging human knowledge and assistance

## Environment Shaping

Process of applying modest **modifications** to the current environment to make it more agent-friendly and minimize the occurrence of negative side effects

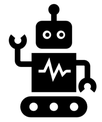
Example:

- Agent needs to push a box $B$ from location $L_1$ to $L_2$ as quickly as possible
- Pushing the box over the rug dirties it as a side effect — unknown to agent
- User can assist by covering the rug with a protective sheet




## Actor-Designer Framework

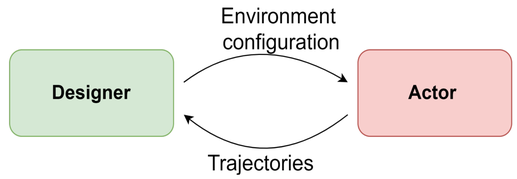
Actor



Optimize assigned task

Designer

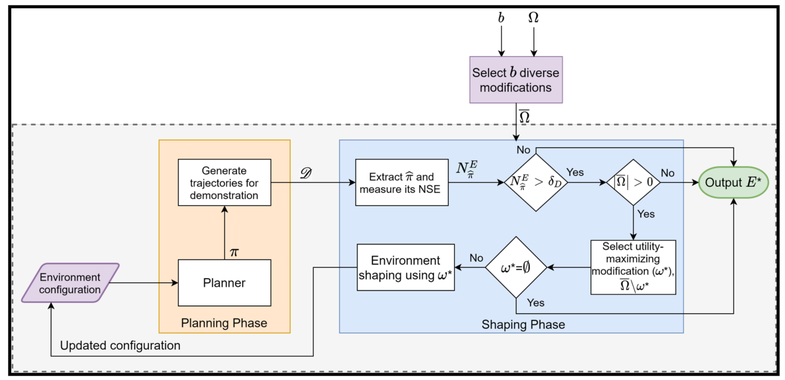
Minimize agent's negative side effects




Designer's slack ($\delta_D$): Tolerance of negative side effects

Actor's slack ($\delta_A$): Maximum deviation from optimal value in the initial environment

- Designer observes trajectories and knows agent's objective, but does not know actor's model
- Actor does not have knowledge of designer's model
- Modifications are applied tentatively for evaluation and reset if they do not produce desired behavior



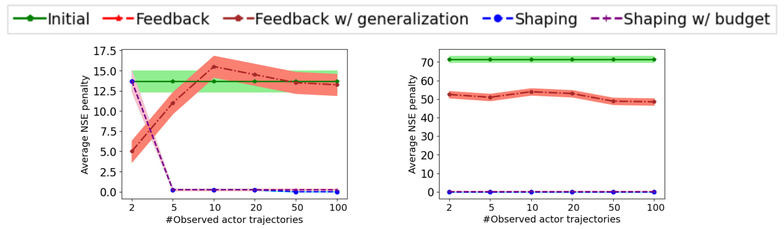




Figure: Results on boxpushing domain: avoidable side effects (left) and unavoidable side effects (right)

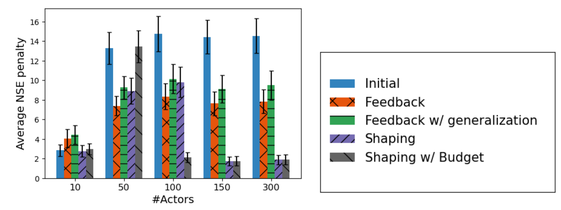



Figure: Driving domain results with multiple actors

## Discussion and Future Directions

- Our proposed framework is effective in mitigating undesirable side effects
- Our user study results shows users are willing to engage in environment shaping
- In the future, automatically identify valid modifications for a problem

Acknowledgments: Support for this work was provided in part by the Semiconductor Research Corporation under grant #2906.001.