

# Autonomous Generation of Robust and Focused Explanations for Robot Policies

**Oliver Struckmeier**, Mattia Racca and Ville Kyrki

[oliver.struckmeier@aalto.fi](mailto:oliver.struckmeier@aalto.fi)

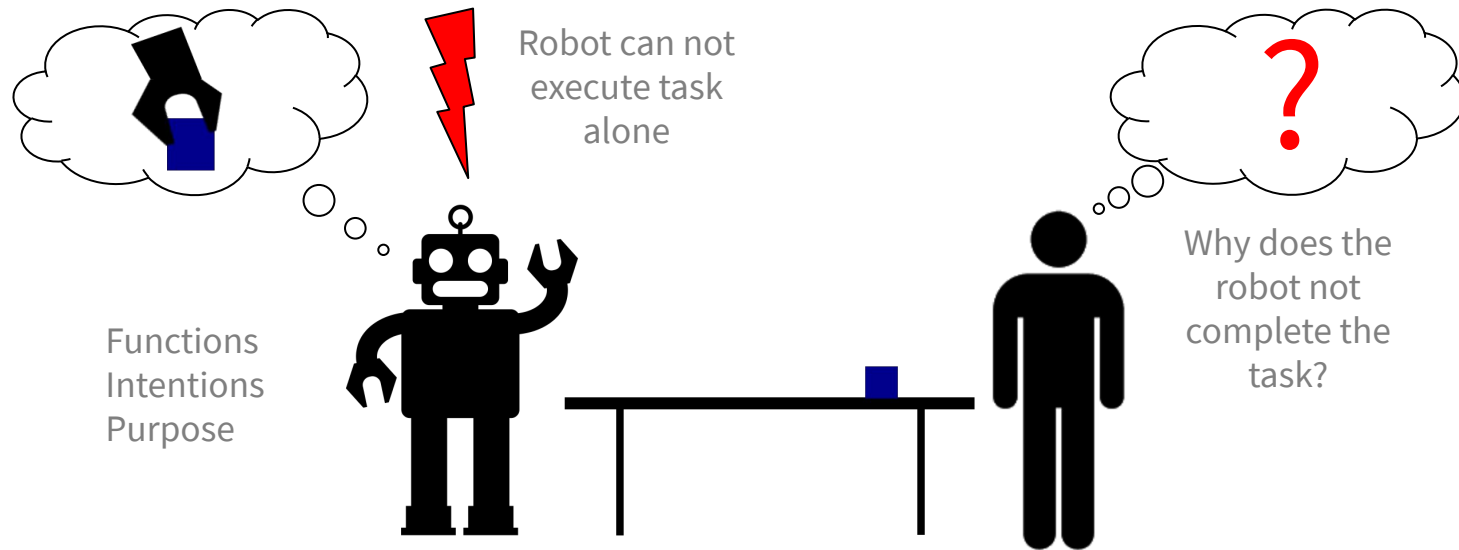


IEEE International Conference on Robot & Human  
Interactive Communication (RO-MAN), 2019

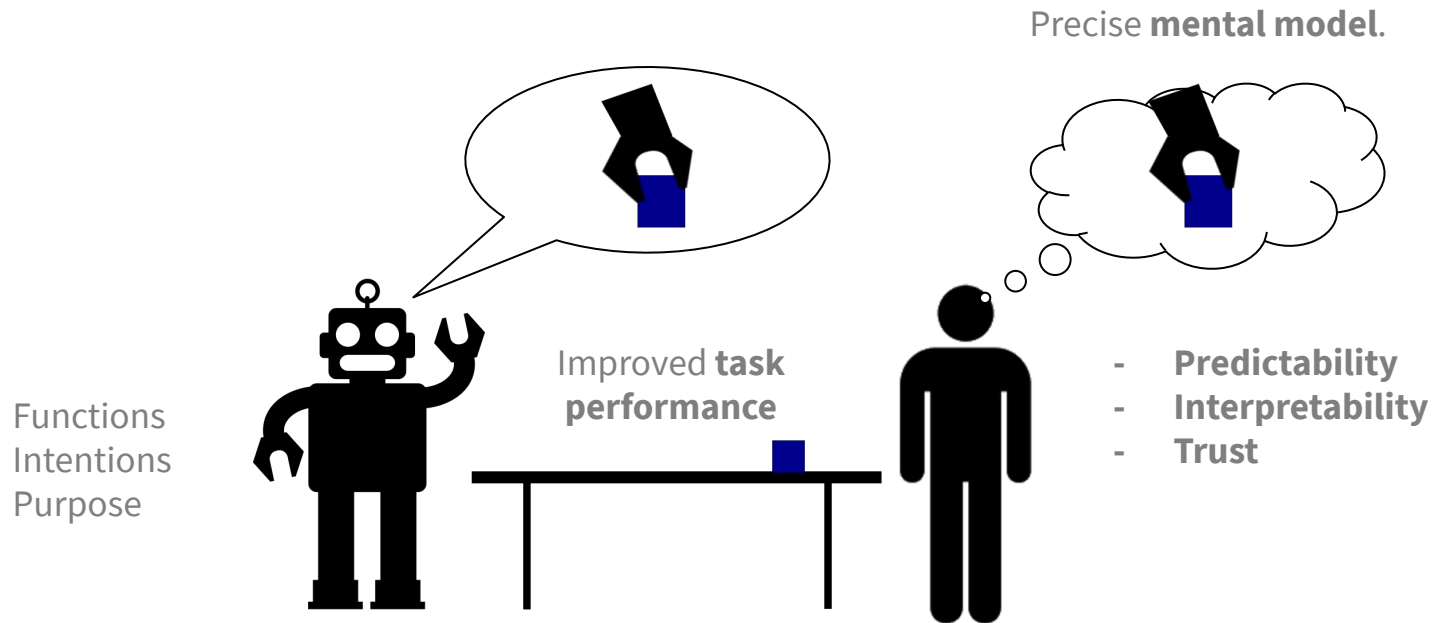


**Aalto University**  
**School of Electrical**  
**Engineering**

# Why should robot's behaviors be transparent?



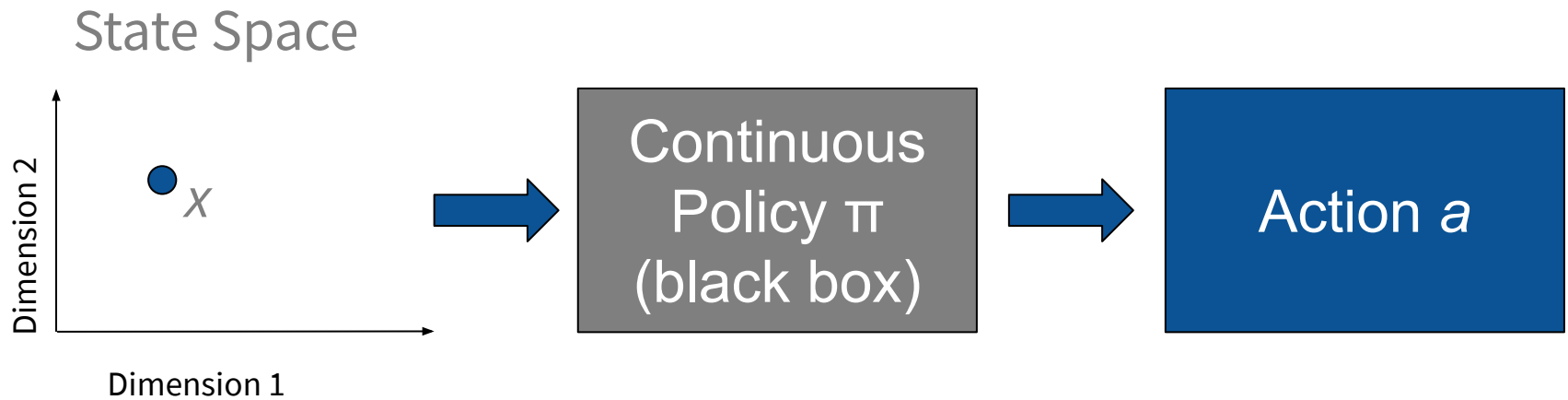
# Why should robot's behaviors be transparent?



How can a robot  
**communicate** its  
decision-making to a  
human user?

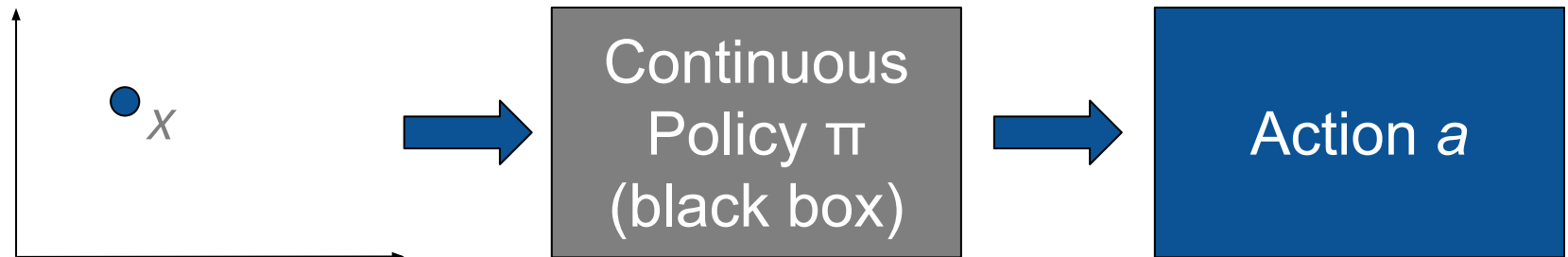
**Explanations!**

# Explaining a robots policy (decision making)



# Explaining a robots policy

State Space



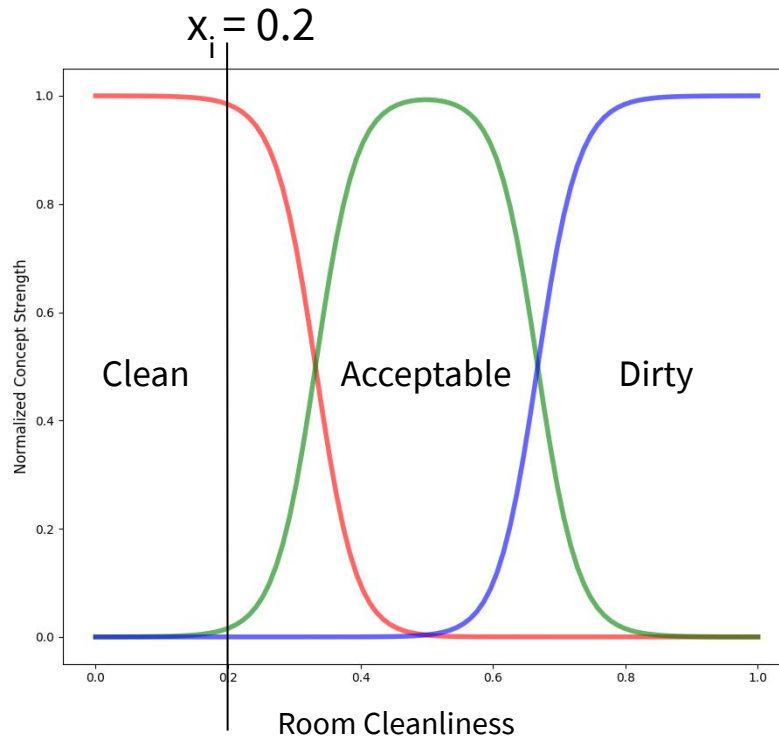
# Structuring explanations

Different ways of explaining policies:

- **Why** did the system do X?

I did **action  $a$**  because **dimension  $d_1$**  was  **$\gamma_1$**  and  
**dimension  $d_2$**  was  **$\gamma_2$**  and  
...

# Making continuous state space dimension human understandable

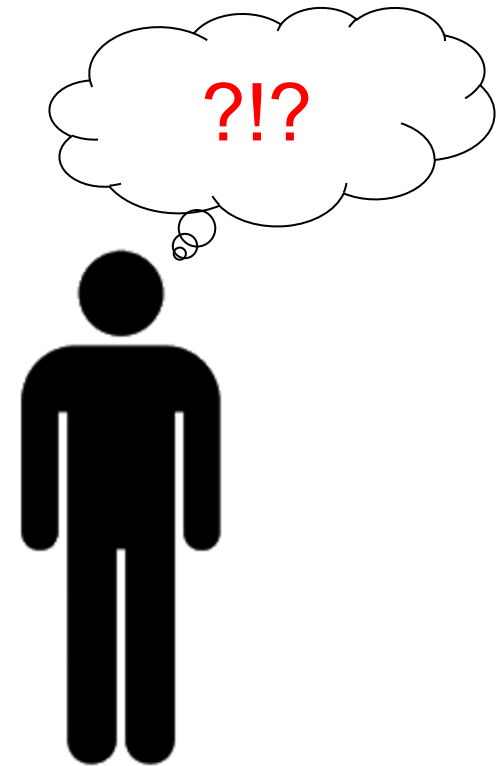
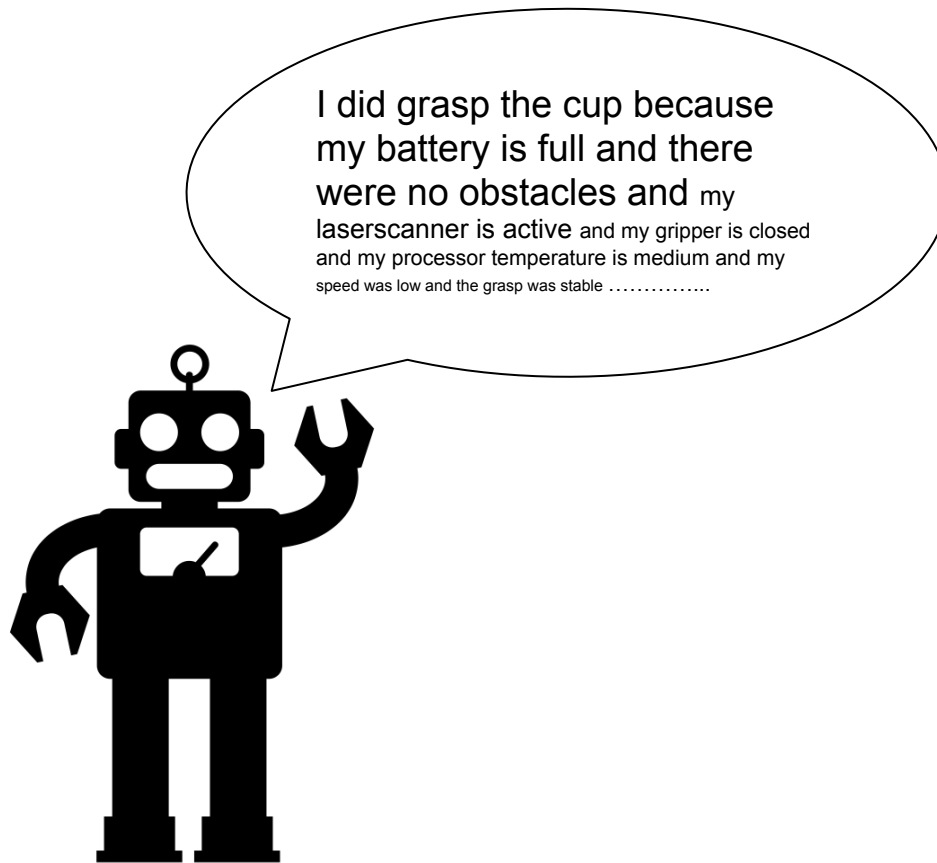


$$\gamma_i(x_i) = \text{“Clean”}$$

**Fuzzy membership function** to assign a natural language descriptor  $\gamma_i$  to a value in a given dimension  $i$ .

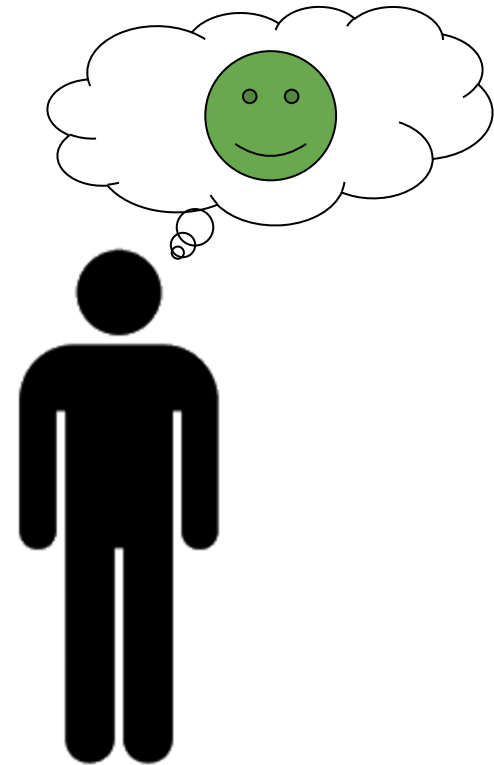
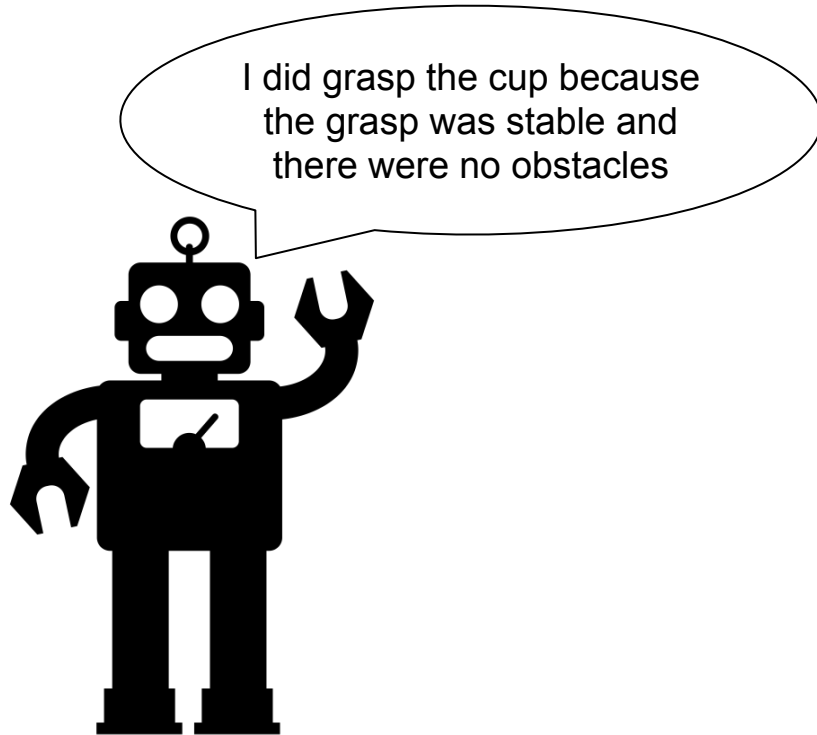


# Explain everything?



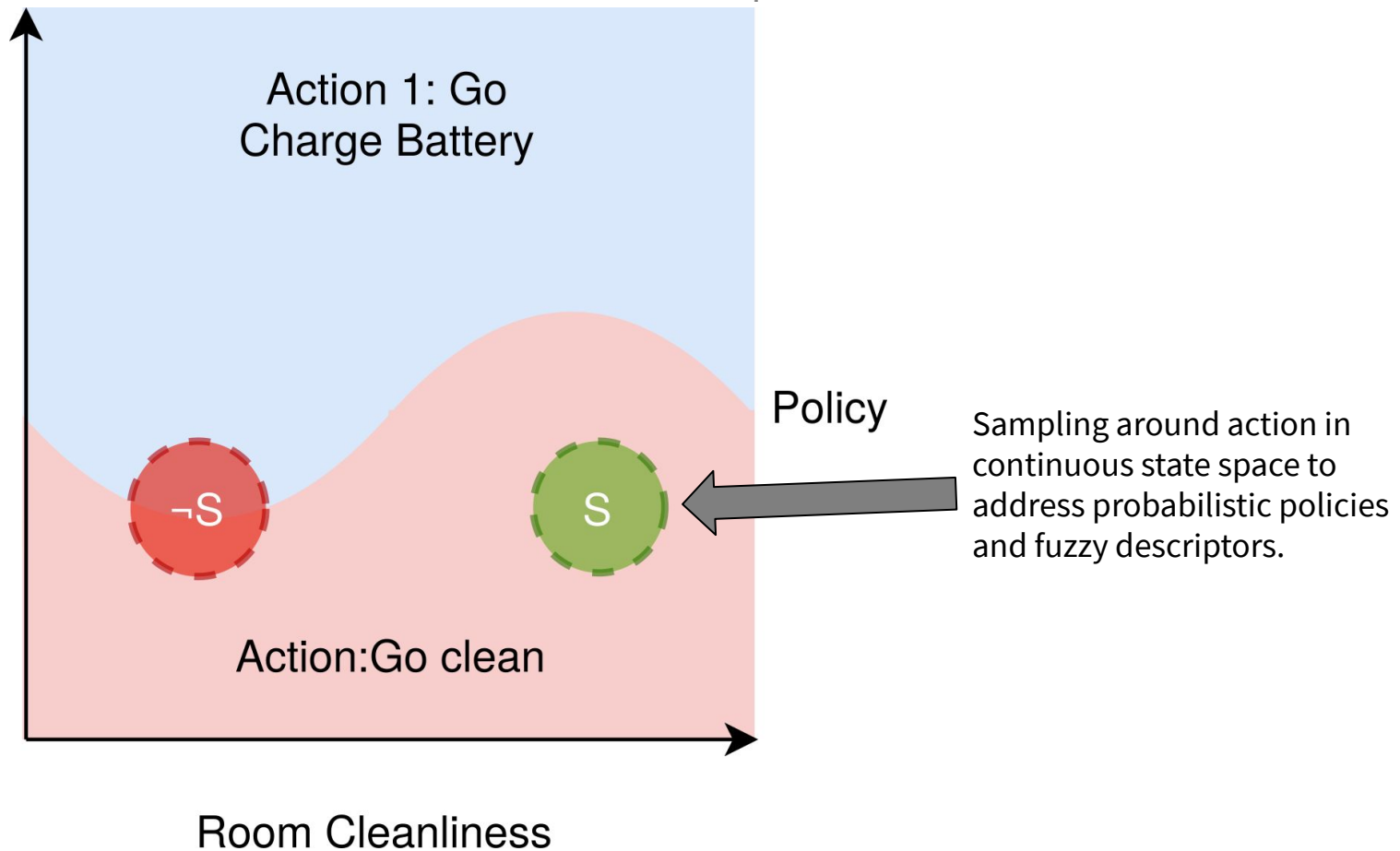
# Focusing explanations on what matters for the decision making

When humans explain the policy of a system, they focus on the most important variables.



# Focusing explanations on what matters

Measure 1) **Action stability**  $S_i$  for dimension  $i$



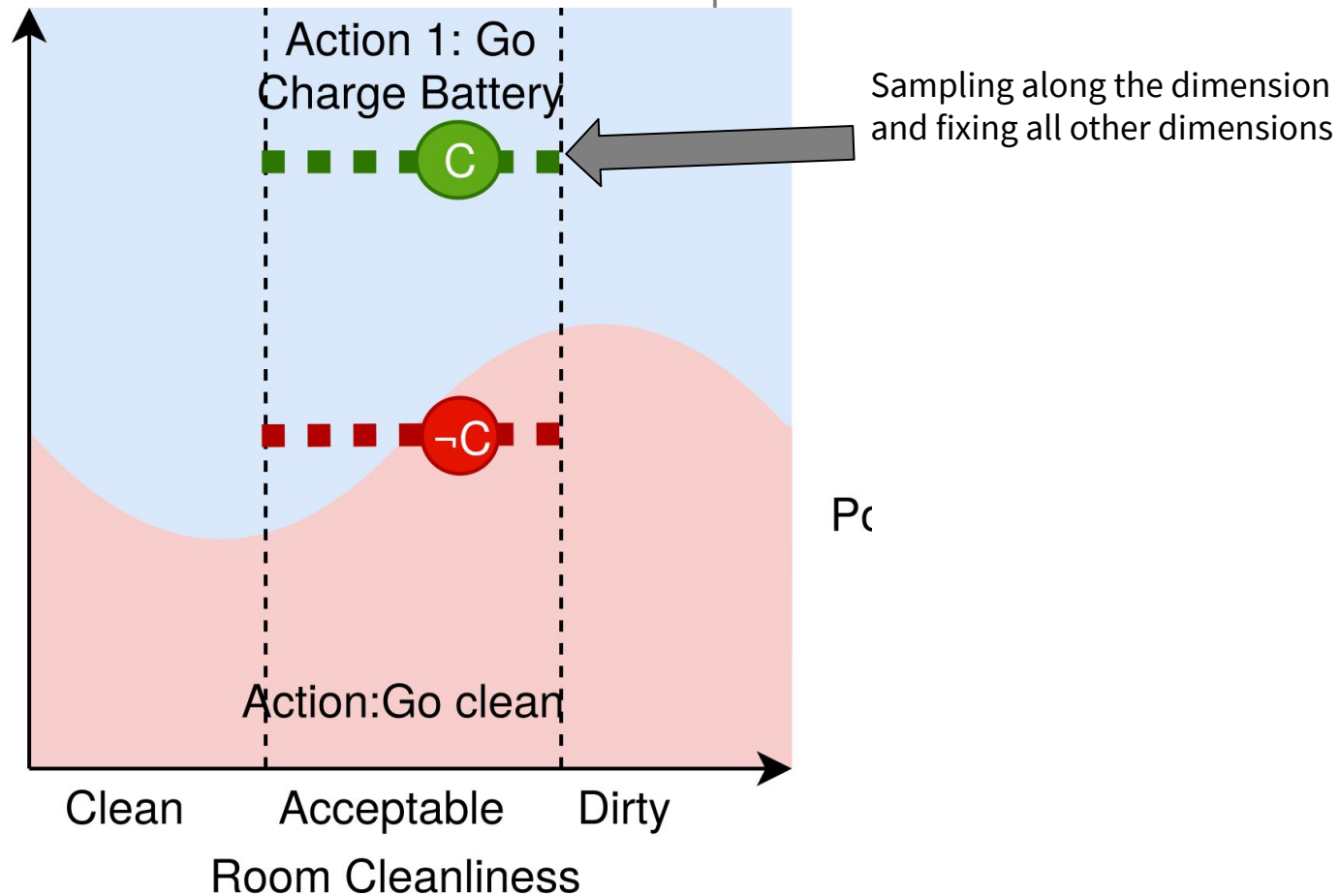
# Focusing explanations on what matters

Measure 1) **Describability**  $D_i$



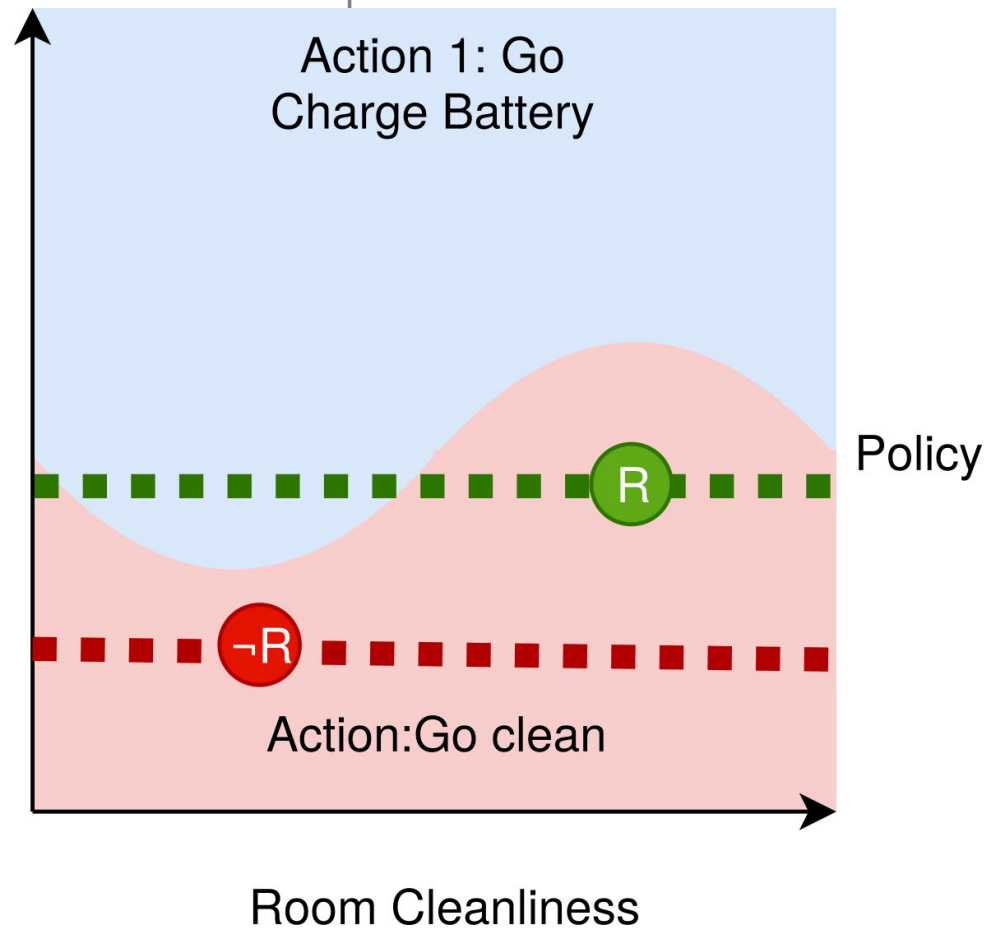
# Focusing explanations on what matters

Measure 3) **Consistency**  $C_i$  for dimension  $i$



# Focusing explanations on what matters

Measure 4) **Relevance**  $R_i$  for dimension  $i$



# Focusing explanations on what matters

Compute one measure for how good a dimension is to explain the action:

$$Q_i = S_i \square D_i \square C_i \square R_i$$

ACTION STABILITY

CONSISTENCY

DESCRIBABILITY

RELEVANCE

# What about real users?

User study: 18 participants.

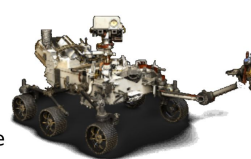
## Hypotheses:

- Users have better policy understanding with F as opposed to C.
- Users prefer the shorter explanations of F over C.
- Users can better detect which parameters matter for action selection when using F.

**1** Mars Rover - learning phase

Battery Level	Ground Quality	Signal Strength	Storage	Temperature
low middle high 92.0	low high 0.31	low middle high 0.83	low high 25.5	low middle high 9.3

**2** Possible Actions:  
1: move  
2: stop and charge  
3: stop and collect ground sample  
4: send data to earth  
5: return and unload the collected samples



Current Scenario: 2/3  
State: 1/7  
Next State

**3** Explanations:  
The action was: move

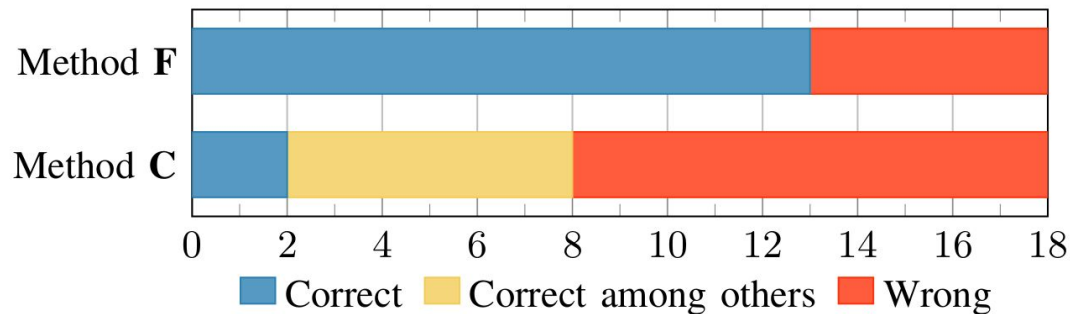
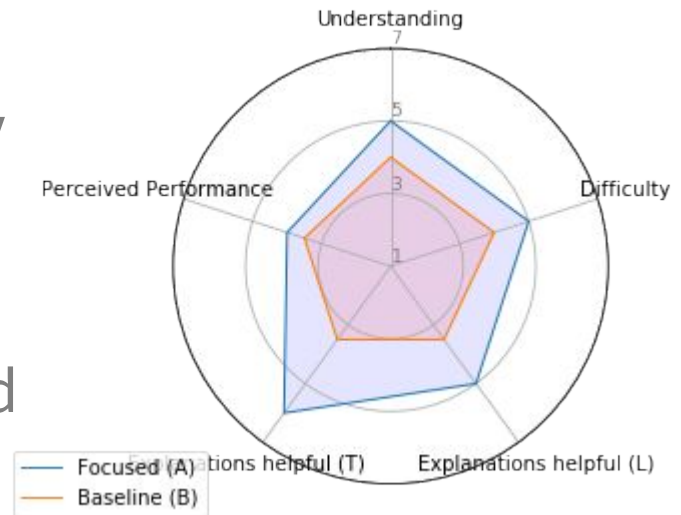
**"I did move because Battery Level was high and Ground Quality was low"**

**Experiment Interface** with focused explaining using the 2 dimensions with the highest value  $Q$



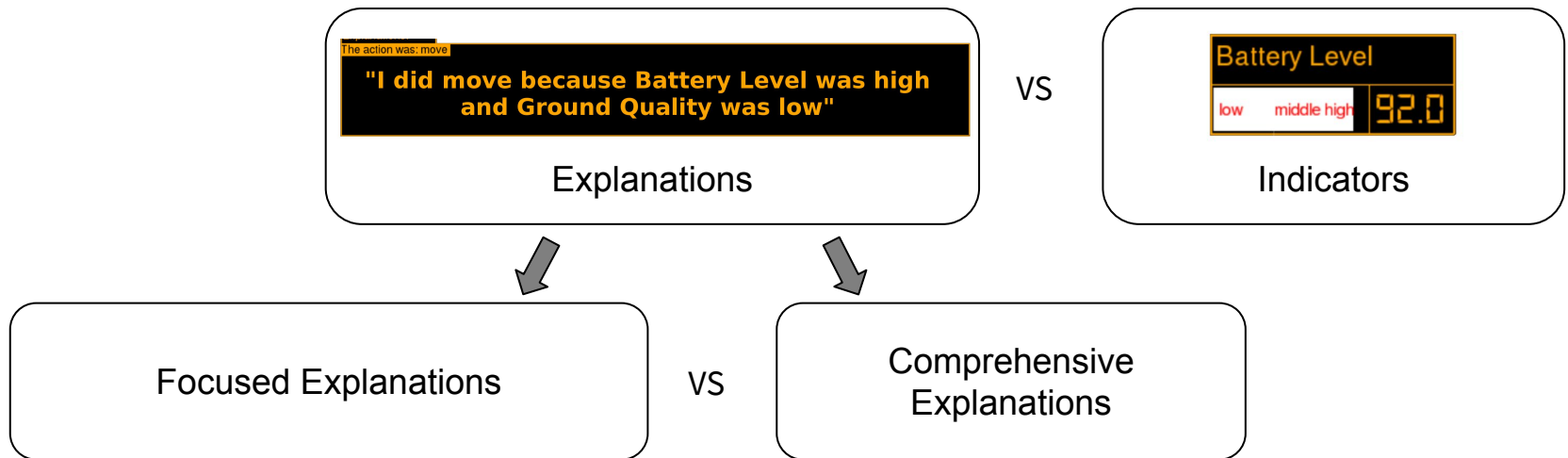
# Performance

- On average, explanations provided by our **Focused method were rated higher**
- No significant differences in measured policy understanding (% of correct state action pairs during testing)
- **Ability to identify irrelevant dimensions better with focused explanations**



# Discussion

=> Explanations depend on the individual preferences



=> How to give explanations if the agent can't explain?

=> Users requested more semantic information:

“Why does a certain parameter value matter for the selected action?”

# Autonomous Generation of Robust and Focused Explanations for Robot Policies

**Oliver Struckmeier**, Mattia Racca and Ville Kyrki

oliver.struckmeier@aalto.fi

**Thank you for the attention!**

Code available at:

**[github.com/Oleffa/FocusedPolicyExplanation](https://github.com/Oleffa/FocusedPolicyExplanation)**



**Aalto University  
School of Electrical  
Engineering**