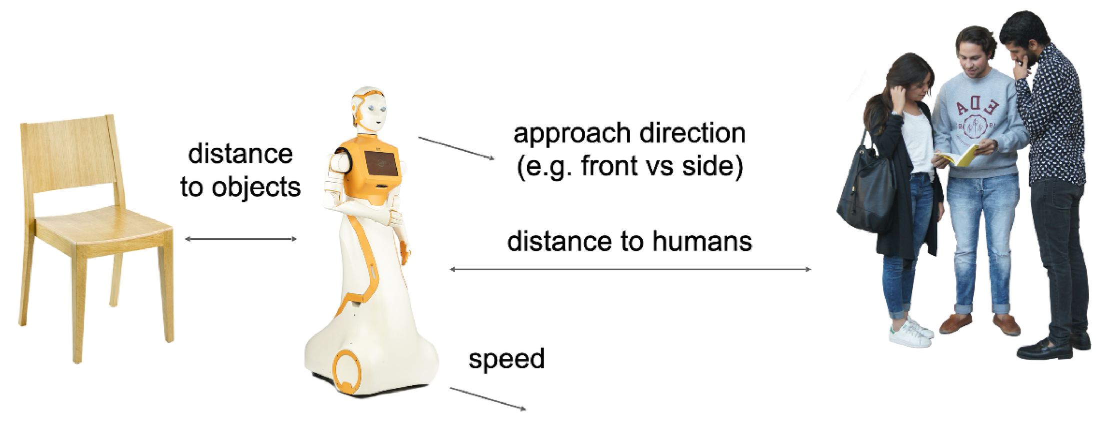
PREFERENCE-BASED LEARNING FOR SOCIAL ROBOTICS

Anand Ballou, Chris Reinke, Xavi Alameda-Pineda
Centre Inria de l'Université Grenoble Alpes, Grenoble, France

Winter School, Grenoble Feb 19th-23rd, 2024
Multimedia Intelligence Artificial Robotics Social

Motivation

We address the problem of behavior generation in a social context. We want our robot to handle different scenarios such as group navigation, under various conditions with different success metrics.

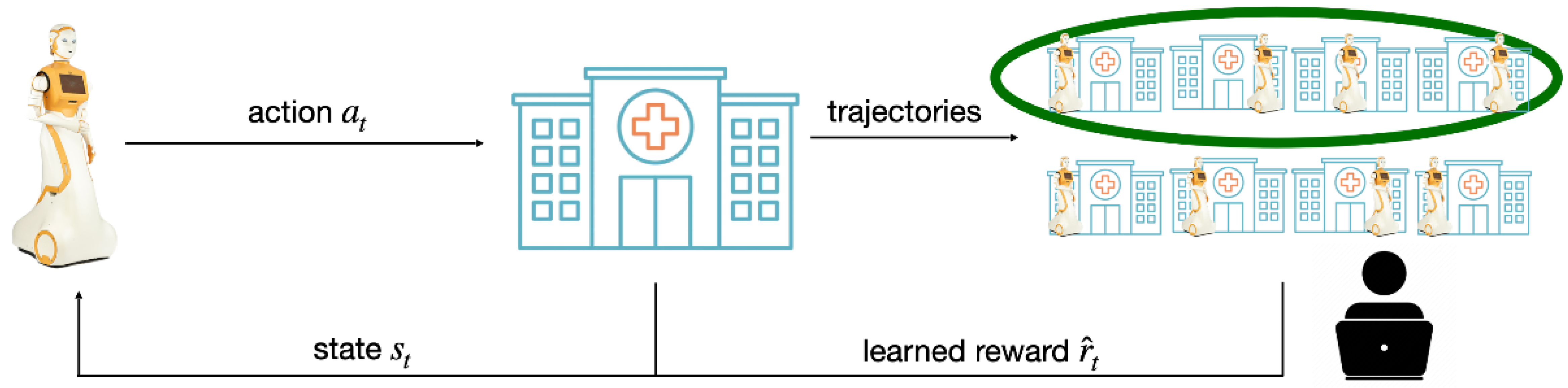


Contribution

A two-step approach, which separates policy training from feedback aggregation, is features agnostic and requires little feedback.

Preference based Learning

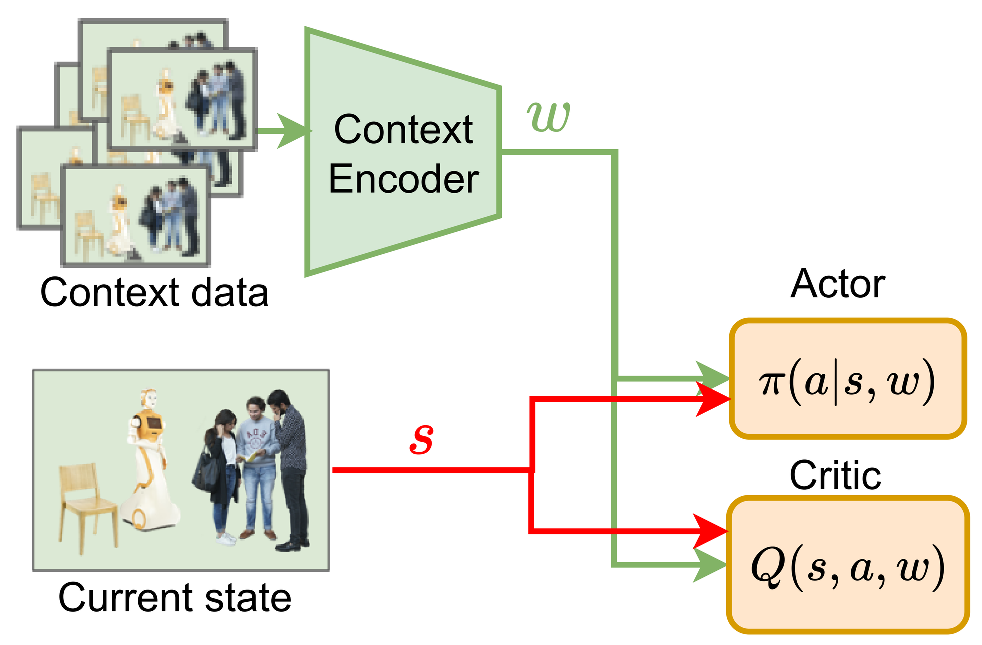
- Add human feedback into the agent learning loop
- Teach agent by expressing preferences about its behaviour



Two drawbacks:
- Requires availability of human at all time
- Requires a high number of feedback or dependent on reward features

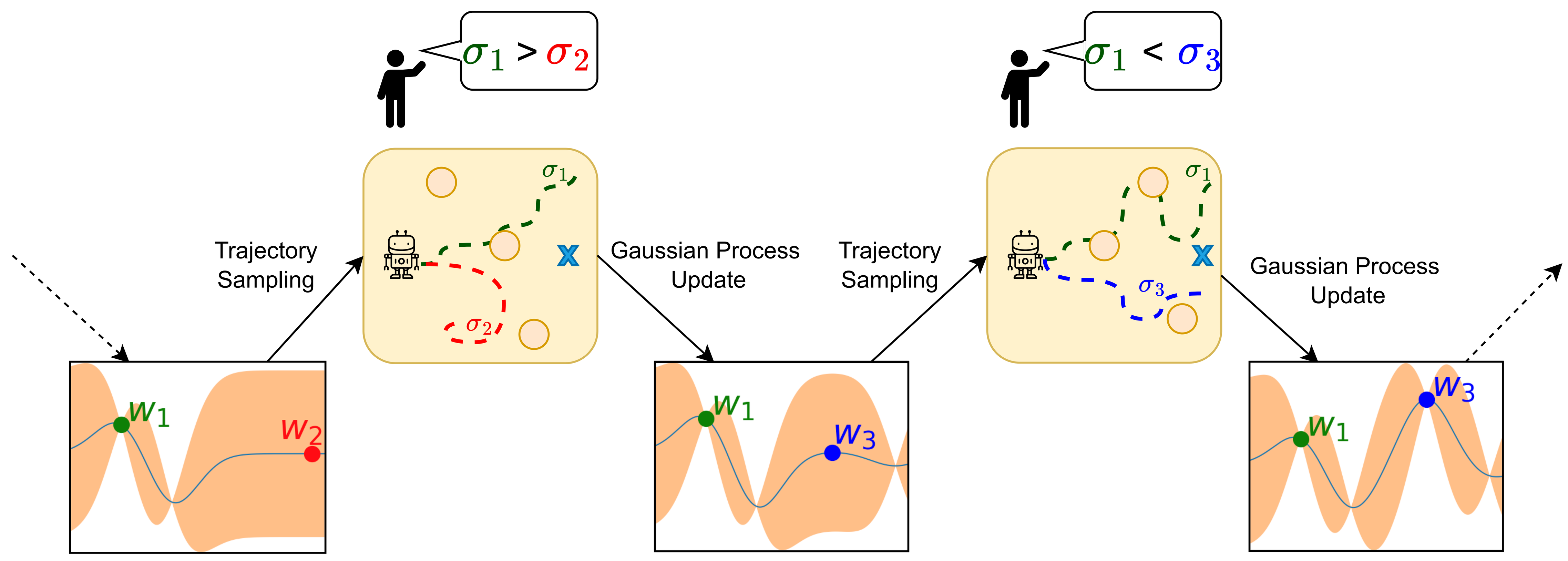
Phase 1: Meta-RL

- A stochastic encoder (VAE) is learned during the training process
- Train by adapting to different tasks with random reward function
- Then search for the most fitting policy in the latent space learned by the VAE



Phase 2: Bayesian Optimisation

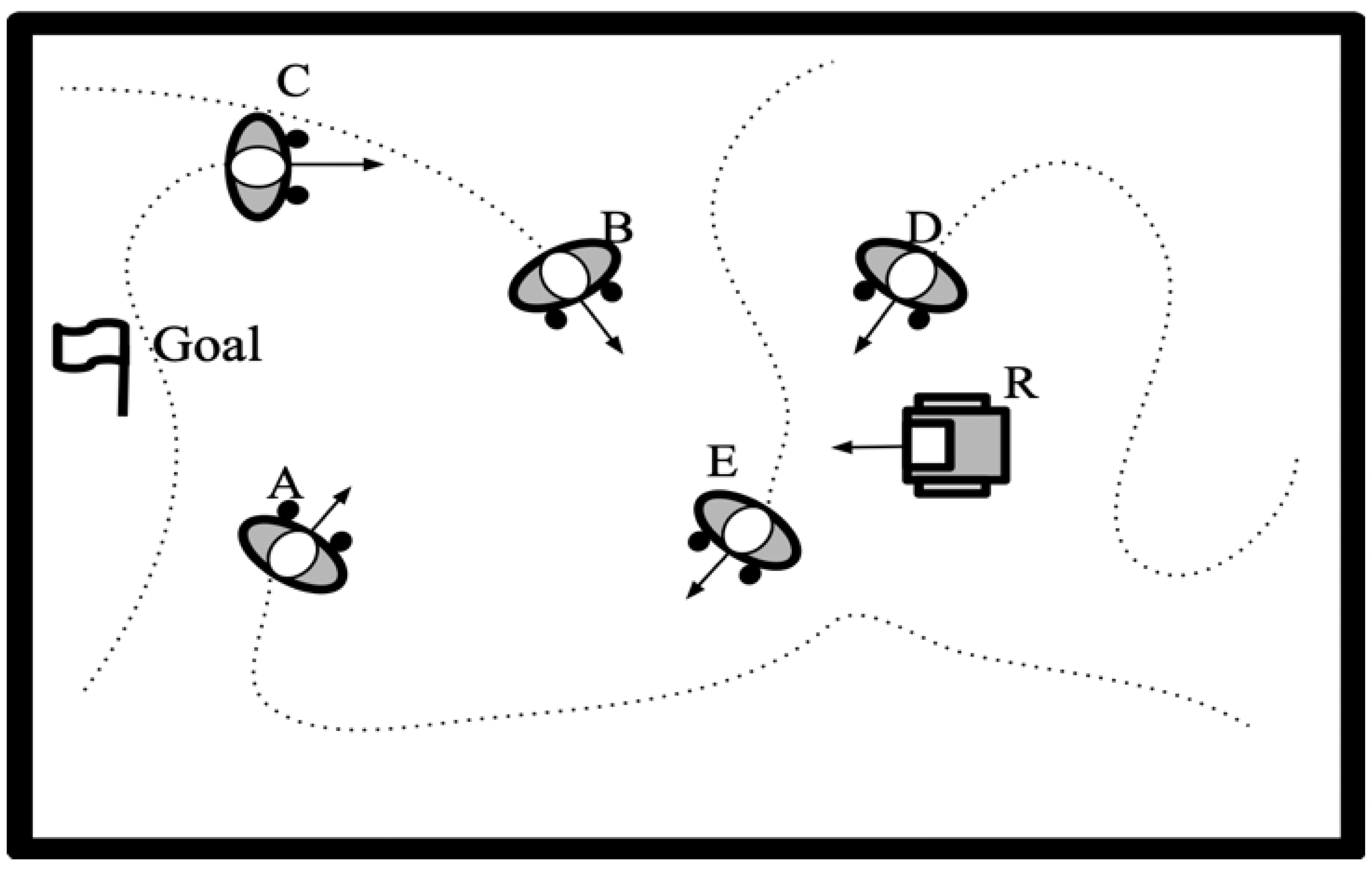
A Bayesian optimization approach is used to select the best policy:



- Sample two policies
- Ask for user's preferences
- Update estimate of our utility function

Experiments

Benchmark on a social navigation environment showing more robust performances with missing reward features and outperforms reward model-based methods in terms of feedback efficiency



Method	No Goal	No Collision	No Social	No Approach	No Speed	Full
Oracle			-31 ± 15			
V-P-BARL			-51 ± 13			
M-P-BARL	-58 ± 16	-48 ± 12	-98 ± 23	-123 ± 40	-175 ± 35	-48 ± 16
Gradient P-BL	-53 ± 8	-53 ± 14	-77 ± 20	-93 ± 21	-145 ± 51	-41 ± 13
Bayesian P-BL	-54 ± 10	-46 ± 13	-105 ± 13	-143 ± 31	-188 ± 43	-38 ± 14

Cumulated reward over different configuration of the environment

Method	No Goal	No Collision	No Social	No Approach	No Speed	Full
V-P-BARL			15 ± 2			
M-P-BARL	18 ± 5	11 ± 4	16 ± 6	9 ± 5	8 ± 6	13 ± 3
Gradient P-BL	73 ± 15	54 ± 10	47 ± 17	53 ± 15	34 ± 11	67 ± 10
Bayesian P-BL	68 ± 12	56 ± 9	49 ± 8	45 ± 12	30 ± 12	56 ± 8

Number of feedback necessary to reach best performances

Summary and Outlook

Investigate the use and limitations of our proposed approach. It presents several advantages:
- No training during feedback collection
- Less feedback required
- Better performances when reward features are ill-defined

Contact information

mail: anand.ballou@inria.fr
team website: team.inria.fr/robotlearn/