Challenges and Directions in Avoiding Negative Side Effects

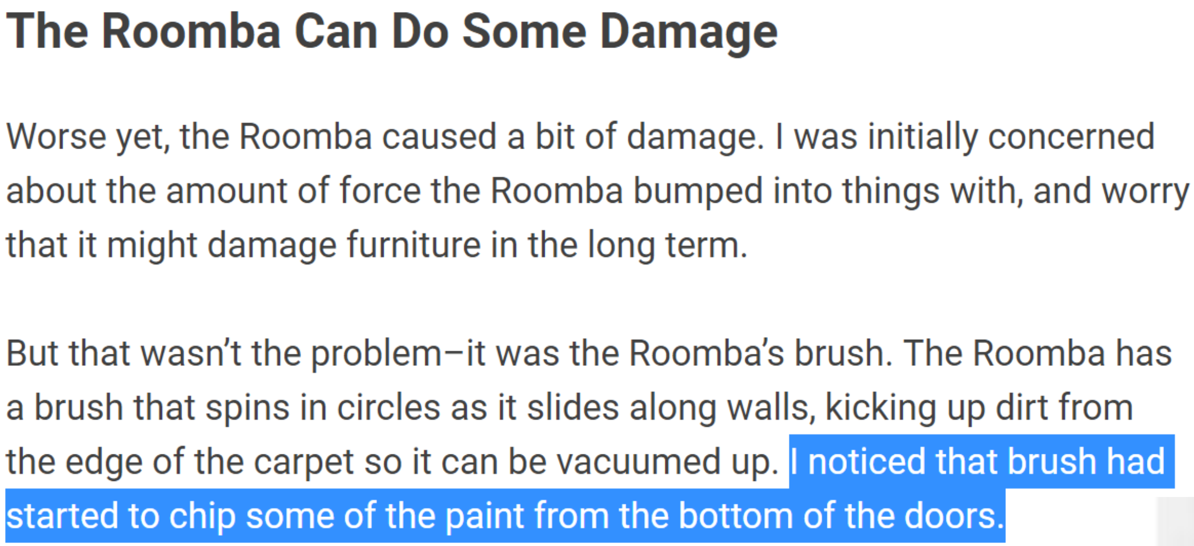
Sandhya Saisubramanian

University of Massachusetts Amherst

AI in the Open World

Unable to avoid puddles – splash water on nearby pedestrians, damage the car

AI in the Open World



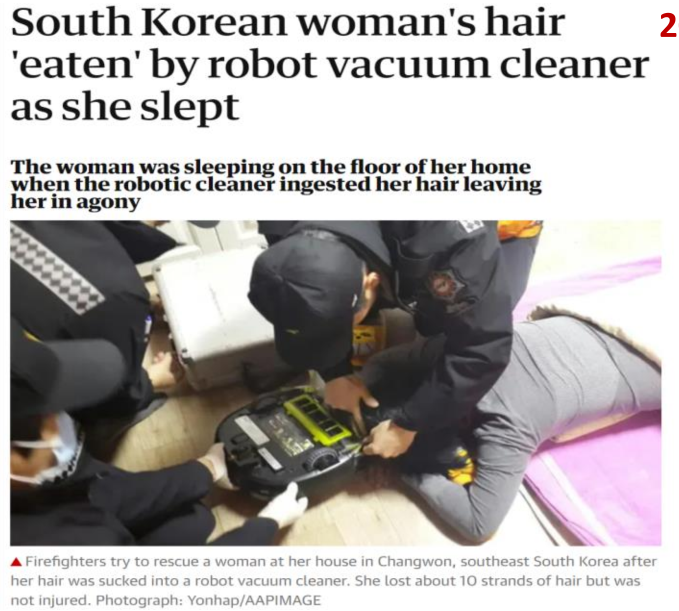
The Roomba Can Do Some Damage

Worse yet, the Roomba caused a bit of damage. I was initially concerned about the amount of force the Roomba bumped into things with, and worry that it might damage furniture in the long term.

But that wasn't the problem–it was the Roomba's brush. The Roomba has a brush that spins in circles as it slides along walls, kicking up dirt from the edge of the carpet so it can be vacuumed up. I noticed that brush had started to chip some of the paint from the bottom of the doors.

AI in the Open World

2015

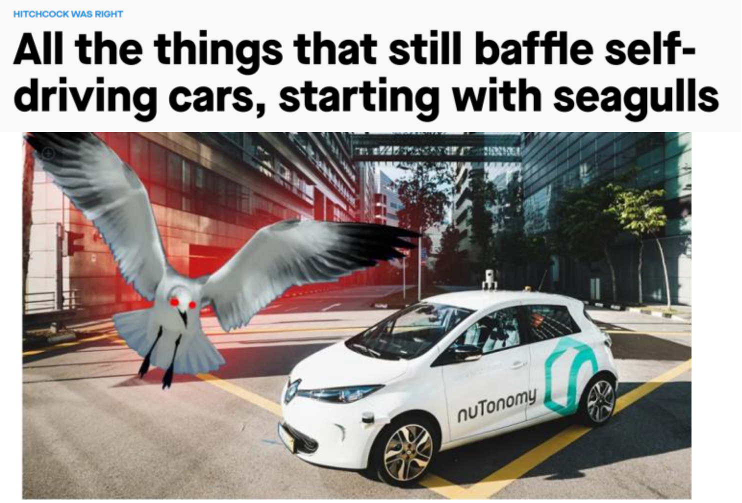


South Korean woman's hair 'eaten' by robot vacuum cleaner as she slept

The woman was sleeping on the floor of her home when the robotic cleaner ingested her hair leaving her in agony

▲ Firefighters try to rescue a woman at her house in Changwon, southeast South Korea after her hair was sucked into a robot vacuum cleaner. She lost about 10 strands of hair but was not injured. Photograph: Yonhap/AAPIMAGE

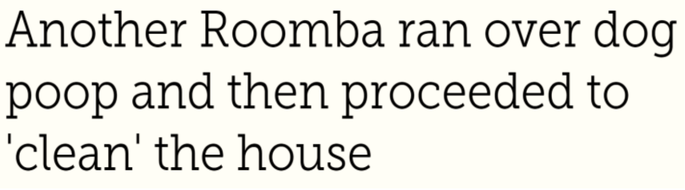
2018



HITCHCOCK WAS RIGHT

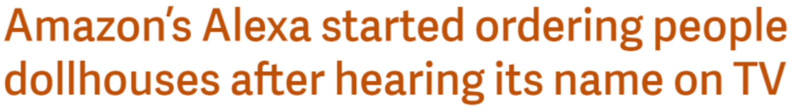
All the things that still baffle self-driving cars, starting with seagulls

2019



Another Roomba ran over dog poop and then proceeded to 'clean' the house

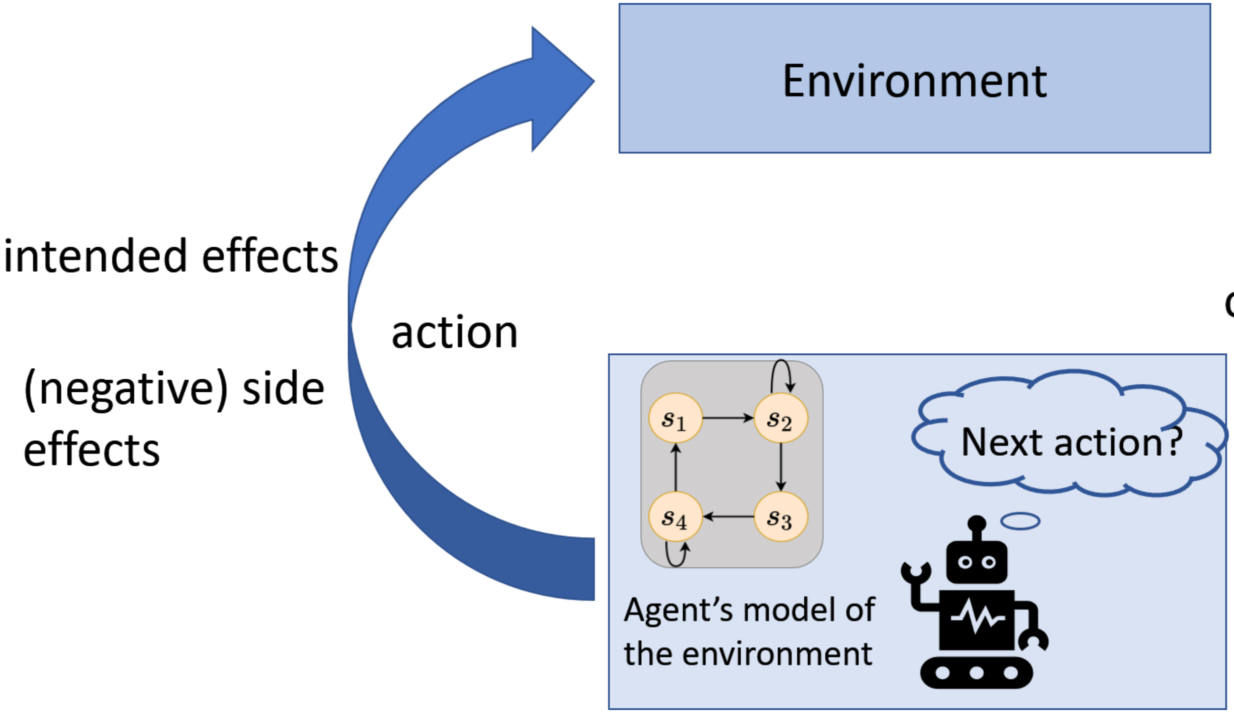
2017



Amazon's Alexa started ordering people dollhouses after hearing its name on TV

Negative Side Effects (NSE)

- Undesired effects of the agent's actions that occur in addition to the intended effects during its operation



Negative Side Effects

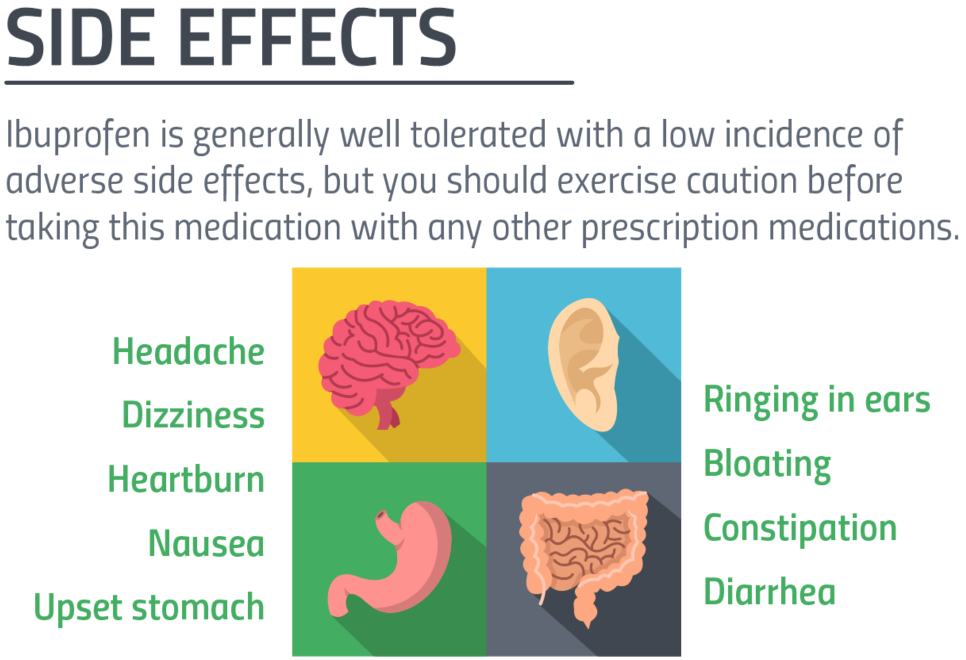
- Affect safety, reliability, and user trust
- Users may stop trusting the system and abandon it
  - even if the system outperforms humans in the task [Dietvorst et al., 2015]
- Inherently challenging to identify during system design

Causes of Negative Side Effects

- Blind to negative side effects
- Objective function focuses on narrow aspects of the environment but its operation affects other aspects
- Incompletely specified models
  - ➢Inadvertently overlooked details
  - ➢Unavailability of accurate information
  - ➢Cultural differences between target users and the development team

Negative Side Effects Occurrence

- System capabilities
- Environment settings
- Assigned task
- User preferences and tolerance



SIDE EFFECTS

Ibuprofen is generally well tolerated with a low incidence of adverse side effects, but you should exercise caution before taking this medication with any other prescription medications.

Taxonomy

Property	Property Values
Severity	Ranges from mild to safety-critical
Reversibility	Reversible or irreversible
Avoidability	Avoidable or unavoidable
Frequency	Common or rare
Stochasticity	Deterministic or probabilistic
Observability	Full, partial, or unobserved
Exclusivity	Prevent task completion or not

Challenges in Minimizing NSE

- Requires broad background knowledge about the environment
- Requires feedback to gather knowledge
- May be irreversible and unavoidable
- Introduce trade-off between assigned task and NSE
- Model revisions are expensive and hard to verify

Recent Approaches

➢ Model and Policy Update

- Update agent’s model and policy based on gathered information
- [Hadfield-Mennel et al., 2017, Saisubramanian et al., 2020]

➢ Constrained Optimization

- Constrain the features in the environment that can be altered by the agent
- [Zhang et al., 2018, Zhang et al., 2020]

➢ Minimizing Deviations from Baseline

- Minimize disruptions to the environment, with respect to a baseline state
- [Shah et al., 2019, Krakovna et al., 2019, Krakovna et al., 2020, Turner et al., 2020]

Recent Approaches

➢ Model and Policy Update

- Update agent's model and policy based on gathered information
- [Hadfield-Mennel et al., 2017, Saisubramanian et al., 2020]

➢ Constrained Optimization

- Constrain the features in the environment that can be altered by the agent
- [Zhang et al., 2018, Zhang et al., 2020]

➢ Minimizing Deviations from Baseline

- Minimize disruptions to the environment, with respect to a baseline state
- [Shah et al., 2019, Krakovna et al., 2019, Krakovna et al., 2020, Turner et al., 2020]

Multi-Objective Approach (IJCAI 2020)

Jointly with Ece Kamar and Shlomo Zilberstein

Distinguished Paper Award

Objective 1:

Optimize for the assigned task

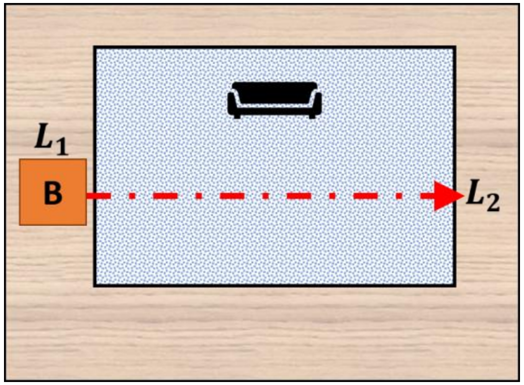
$\succ$

Objective 2:

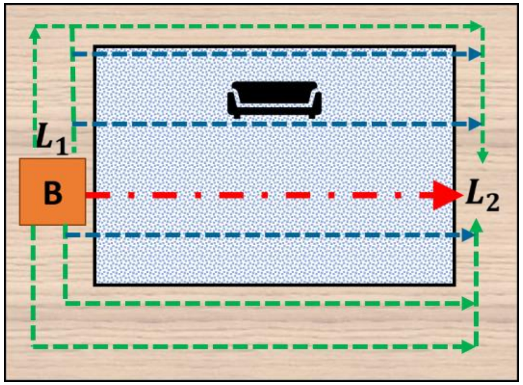
Minimize negative side effects

Maximum allowed deviation (slack): δ

$V_1 = 10$



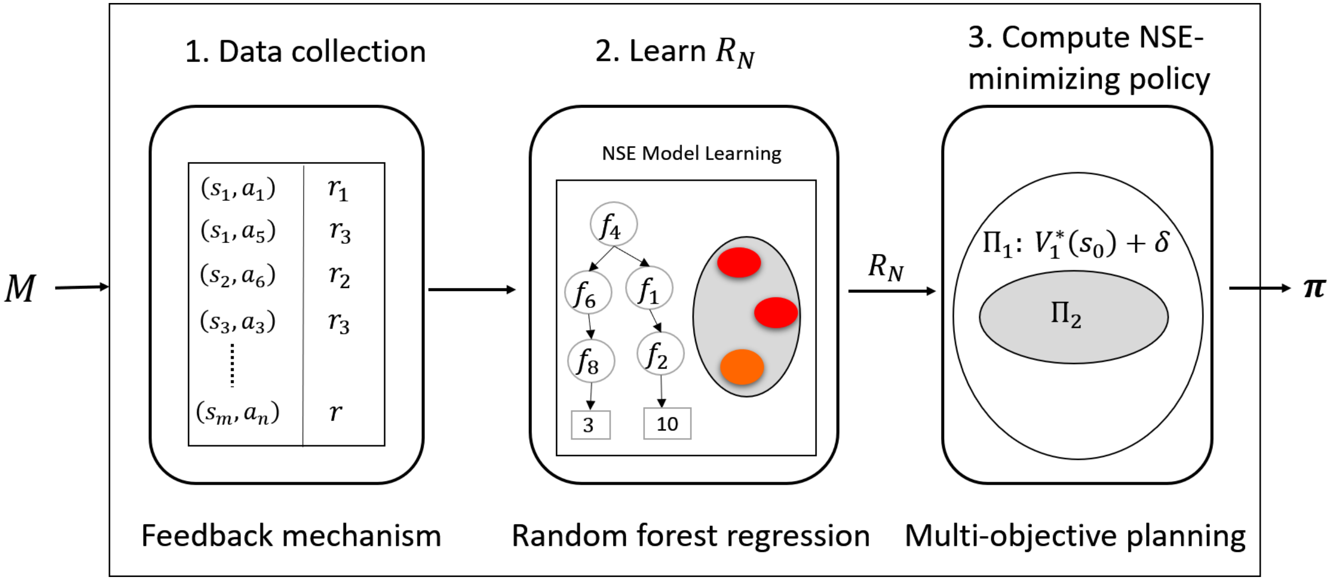
$\delta = 5$ $V_1 \in [10, 15]$



Learning Negative Side Effects

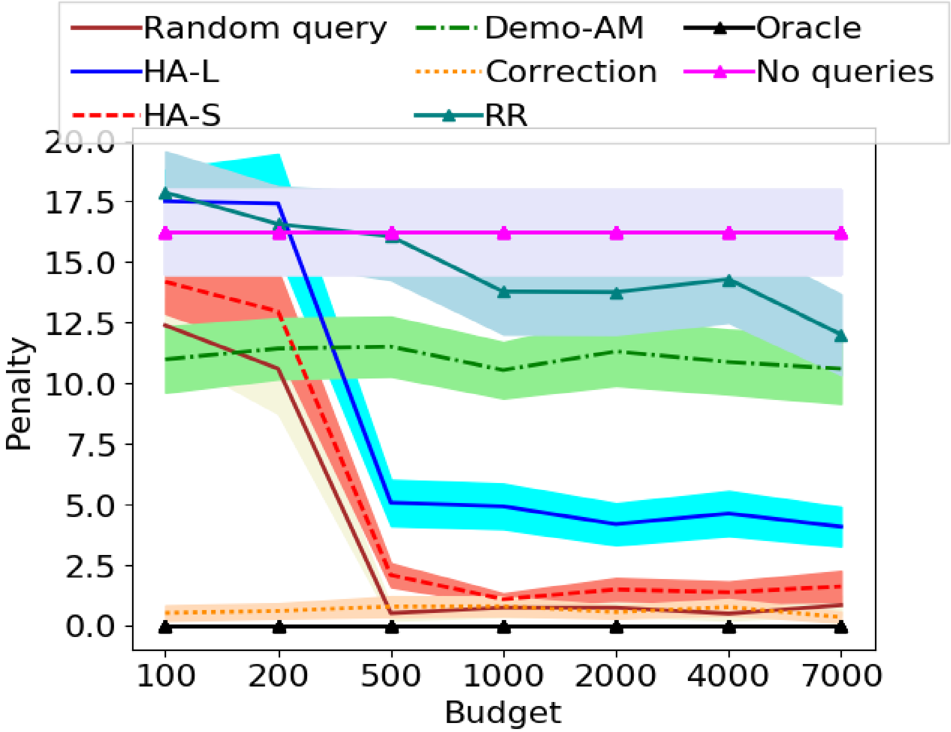
Feedback Mechanisms

- Human approval
- Human corrections
- Human demonstration
- Learning from exploration

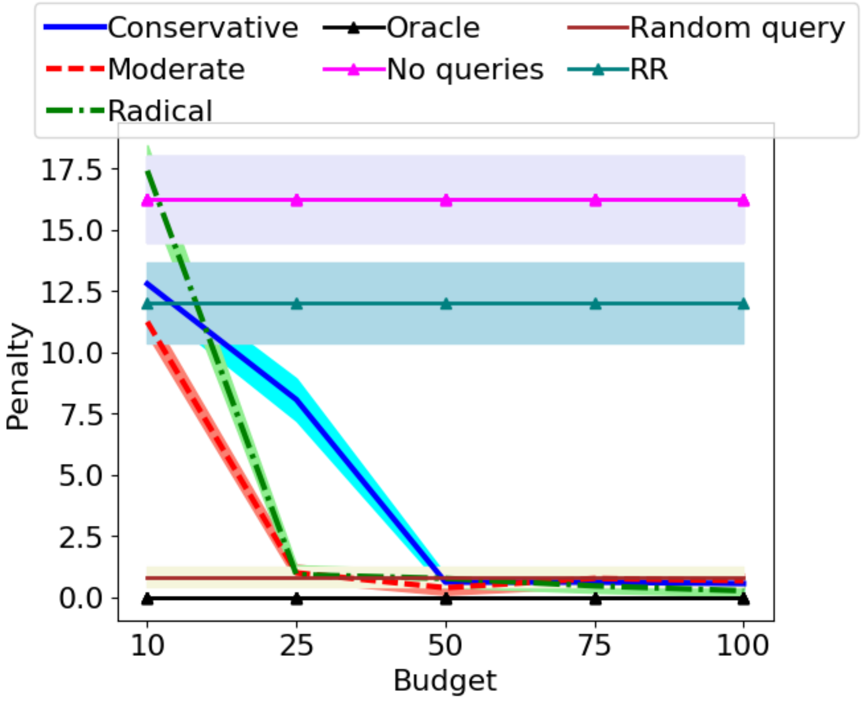


(s_1, a_1)	r_1
(s_1, a_5)	r_3
(s_2, a_6)	r_2
(s_3, a_3)	r_3
⋮	
(s_m, a_n)	r

Experimental Results (more in the paper)



Learning from human feedback



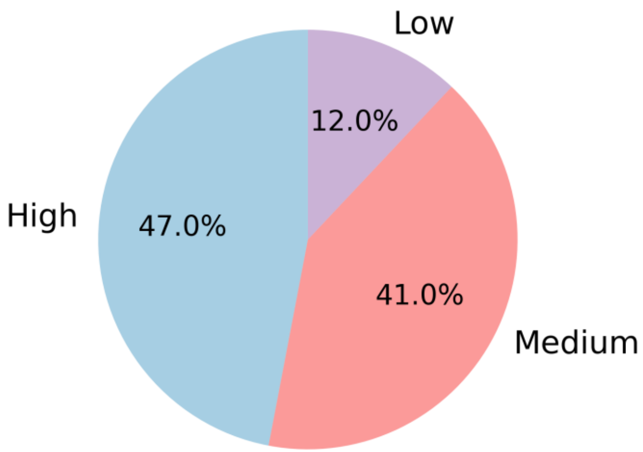
Learning by exploration

User Study

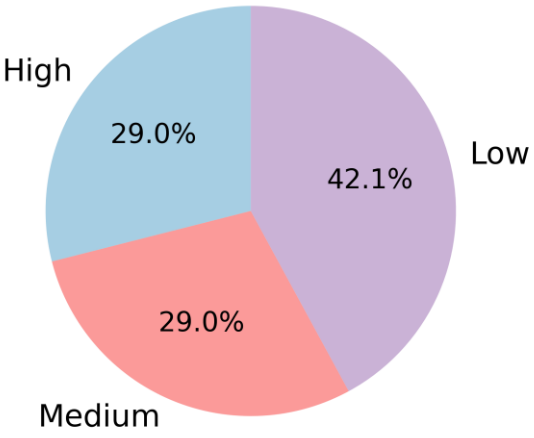
- Study conducted on Amazon Mechanical Turk platform
- 500 participants recruited to complete a pre-survey questionnaire
- 300 participants invited to complete 2 surveys: Roomba and AV
- Roomba side effects: sprays water on the wall when cleaning the floor
- AV side effects: driving fast through potholes (bumpy ride), harsh braking at stop signs (sudden jerks)
- 183 valid responses for each domain

1. Are users willing to tolerate negative side effects that are NOT safety-critical?

- Select tolerance level: low, medium, high
- User tolerance varies and depends on severity of impacts
- Users willing to tolerate mild to moderate impacts but want to minimize it as much as possible



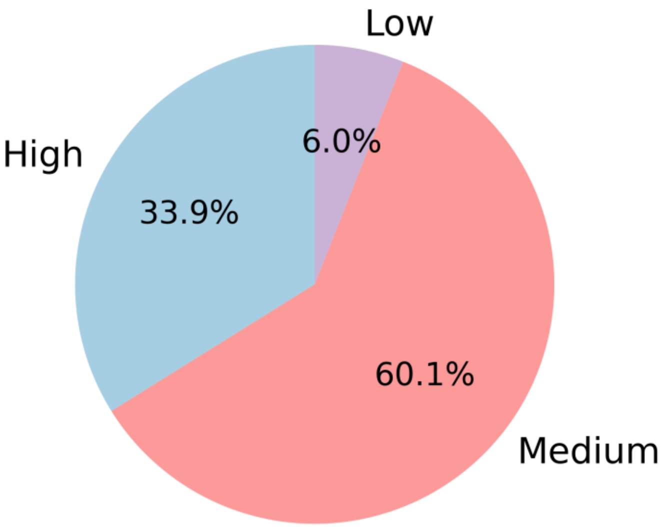
AV Bumpy



AV Sudden Jerks

2. How do negative side effects affect the user's trust in the system?

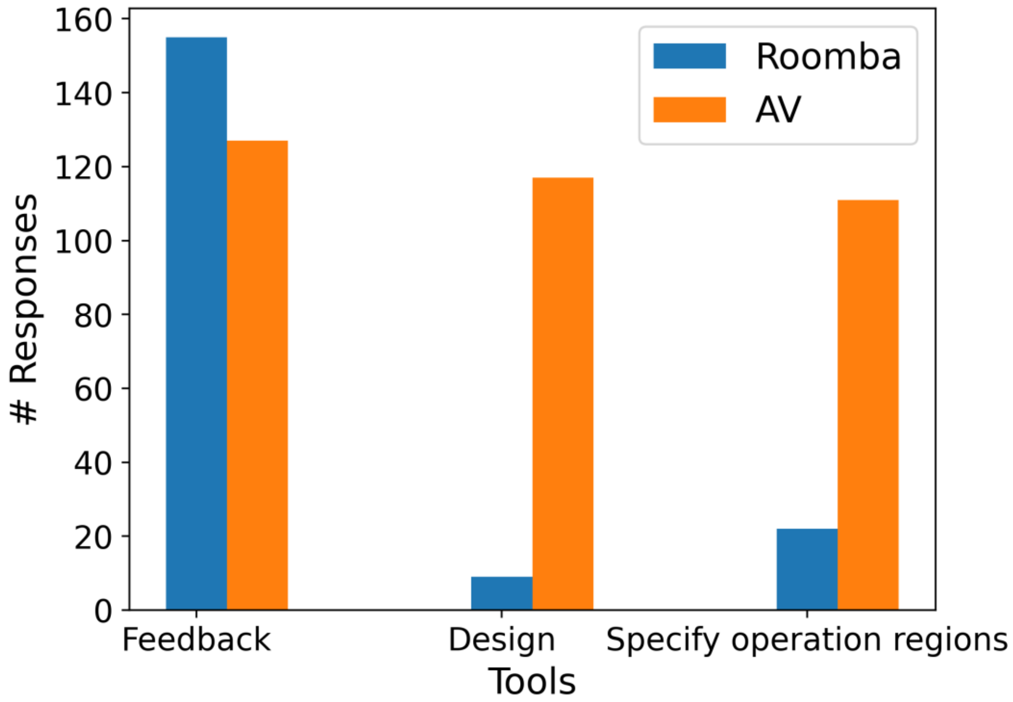
- Low trust: do not trust the system to be capable
- Medium: trust is affected if the system does not adapt over time
- High: trust is unaffected by NSE occurrence



AV Bumpy

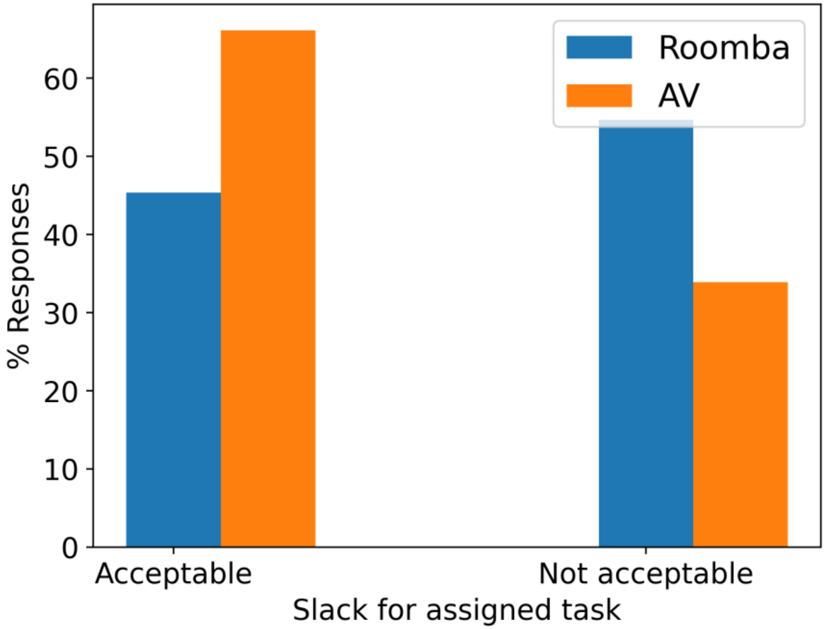
3. Are users willing to assist the system in mitigating the impacts of the side effects?

- Feedback, specifying regions of safe operation, reconfiguring the environment
- Users are willing!
- Feedback – simplest form of interaction with the system



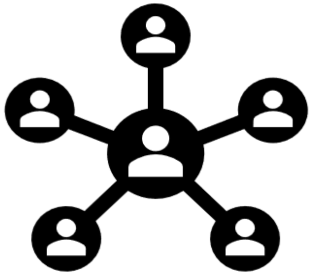
4. Willingness to tolerate a bounded sub-optimal behavior (e.g., taking a longer route)

- Generally willing!
- More positive response for AV since it is similar to human behavior--- taking a longer route to avoid bumpiness

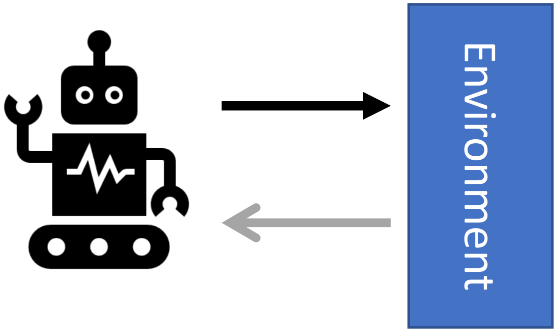


Future Research Directions

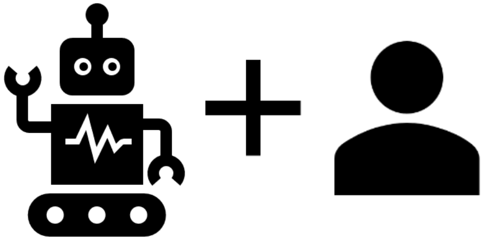
Multi-agent settings



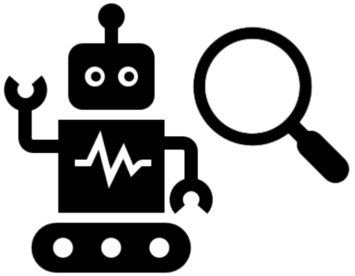
Delayed and noisy feedback



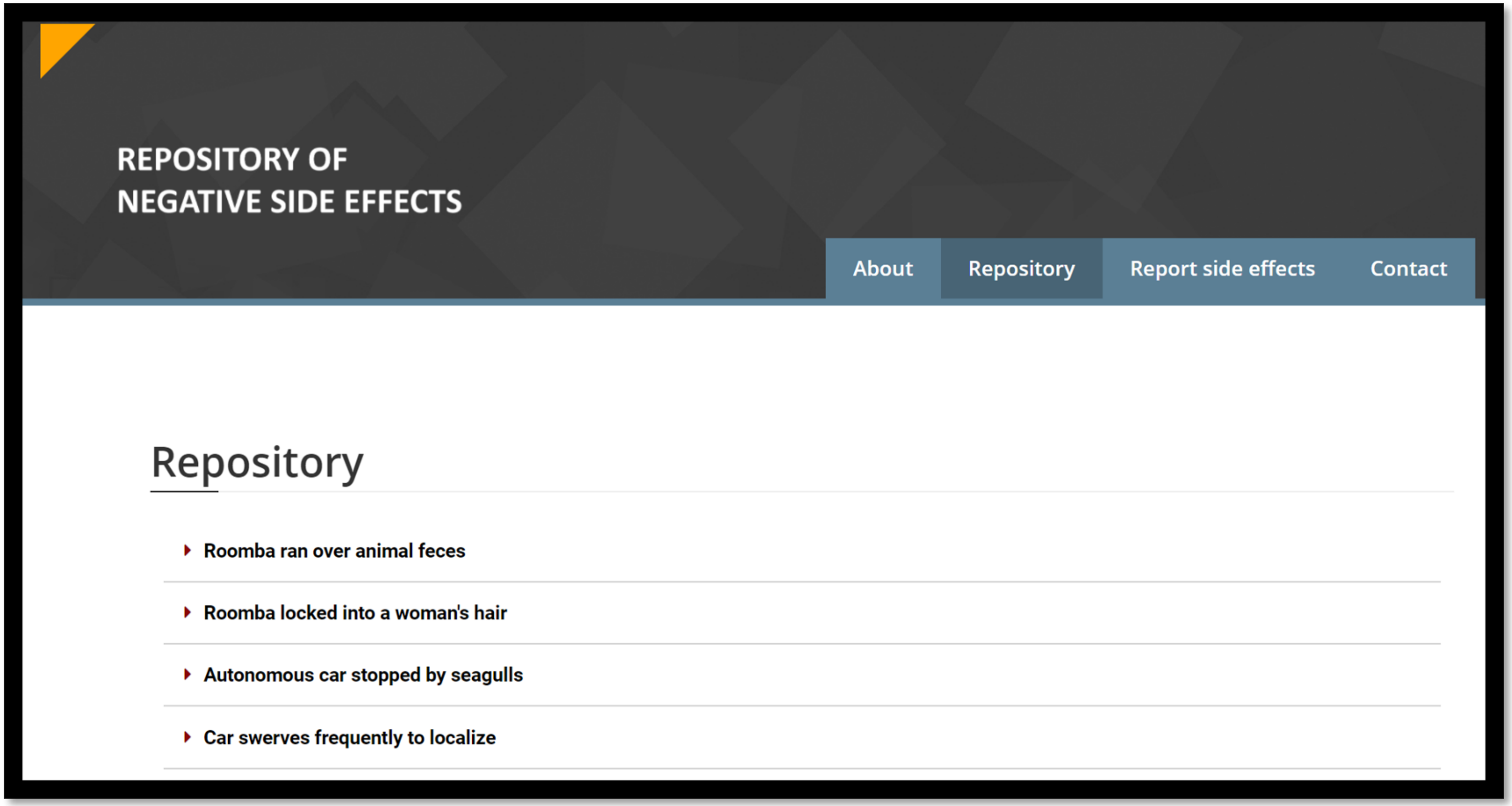
Human-AI teams



Self-monitoring



Public Repository



REPOSITORY OF NEGATIVE SIDE EFFECTS

Repository

- Roomba ran over animal feces
- Roomba locked into a woman's hair
- Autonomous car stopped by seagulls
- Car swerves frequently to localize

https://groups.cs.umass.edu/nse/

Summary

- Negative side effects affect safety & reliability of AI systems
- Negative side effects could affect user trust
- Inherently challenging to detect during system design
- Requires principled approaches and tools to recognize, measure, and avoid negative side effects