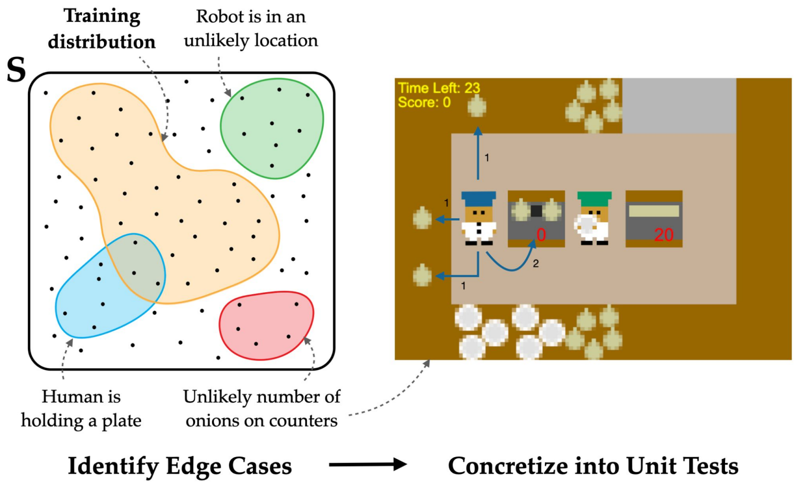
Evaluating the Robustness of Collaborative Agents

Paul Knott, Micah Carroll, Sam Devlin, Kamil Ciosek, Katja Hofmann, Anca Dragan, Rohin Shah

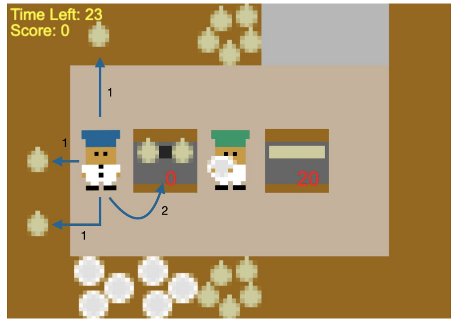
Reward ≠ Robustness

Average test reward is usually used as a proxy for agent robustness. However, using a finite number of evaluation rollouts will rarely cover all edge cases which might make a policy fail. Therefore, *if we cannot rely on test reward, how can we effectively evaluate edge case robustness?*

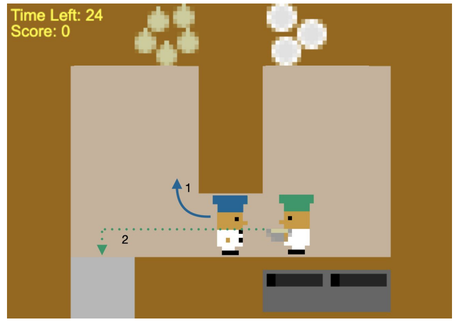


Contributions

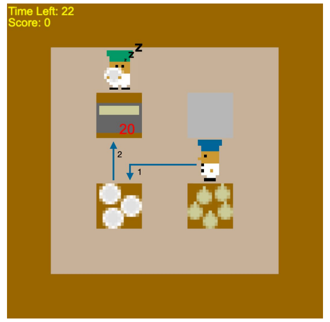
We describe a *methodology for creating* **unit tests to evaluate edge case robustness**. We then create an example suite for the Overcooked environment, and demonstrate that we are able to gain more information about agent robustness with unit tests than through average reward alone.



State robustness test:
multiple onions on counters.



Agent robustness test:
stubborn human partner



Memory robustness test:
Human partner AFK

Robustness Unit Tests

We place the agent in realistic edge cases with respect to the **state and the partner agent**. To create tests, we:

1. Identify qualitative situations in which desired agent behavior is clear;
2. Concretize each situation to a unit test;
3. Improve test coverage by probing the trained agents.

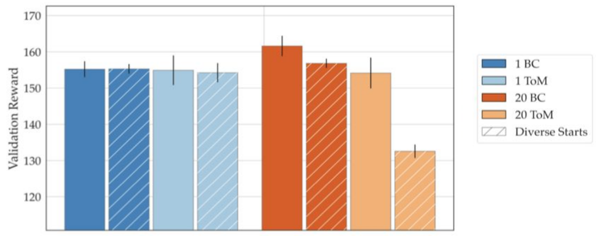
Assessing the utility of robustness tests

To test the effectiveness of our robustness tests, we compare three proposals for improving robustness in human-AI cooperation scenarios:

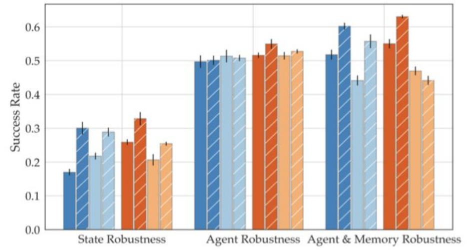
1. Improving the quality of human models (ToM vs BC).
2. Improving the diversity of human models that the agent is trained with
3. Leveraging human-human gameplay data

Our results show that robustness and reward can be relatively uncorrelated:

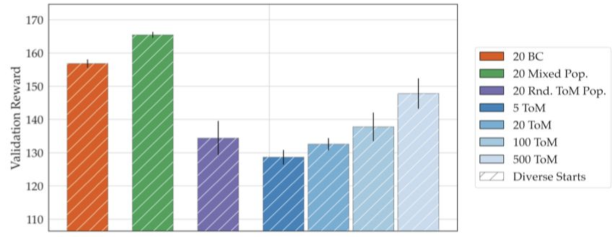
Diverse starts reduces average reward…



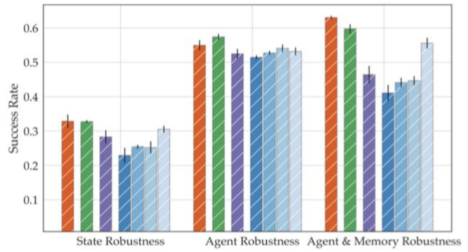
… but improves robustness



Using a mixture of BC and ToM improves reward…



…but tends to keep robustness the same



Robustness guarantees?

While a test suite cannot guarantee full edge case coverage, it is still a significant improvement over the current status quo of only looking at reward – which covers only the edge cases which are randomly encountered.

Moreover, none of the deep RL agents scored above 65% in our robustness tests, suggesting that our approach can serve as a useful metric for the foreseeable future.