Evidence that Robots Trigger a Cheating Detector in Humans

Alexandru Litoiu, Daniel Ullman, Jason Kim, Brian Scassellati
Department of Computer Science, Yale University

Introduction

- Short et al. found that in a game between a human participant and a humanoid robot, the participant will perceive the **robot as more agentic and having more intentionality if it cheats** than if it plays without cheating.
- However, in that design, **the robot that actively cheated also generated more motion than the other conditions**.
- We **disambiguated** between the following two possible causes of the effect:
  - The **additional motion** of the cheating behavior caused greater attributions of agency.
  - A **cheating detector** that has been shown to trigger towards humans also triggered towards the cheating robot, causing greater attributions of agency.
- Our experimental design kept constant the amount of motion while varying the directionality of the cheat from adversarial to pro-social.
- **83 participants** in between-participant design.
- Salience, engagement, and attributions varied as the direction and magnitude of the cheat changed, **supporting the cheating detector hypothesis**.

Experimental Design

Procedure

- Nao robot played 30 rounds of rock-paper-scissors with each participant.
- No cheating occurred in the control rounds.
- The robot would cheat on the first two possible occasions in the cheating rounds, in accordance with the experimental condition.

ROUND #	1	10	20	30
ROUND TYPE	Control Rounds	Cheating Rounds	Control Rounds	

Experimental Conditions

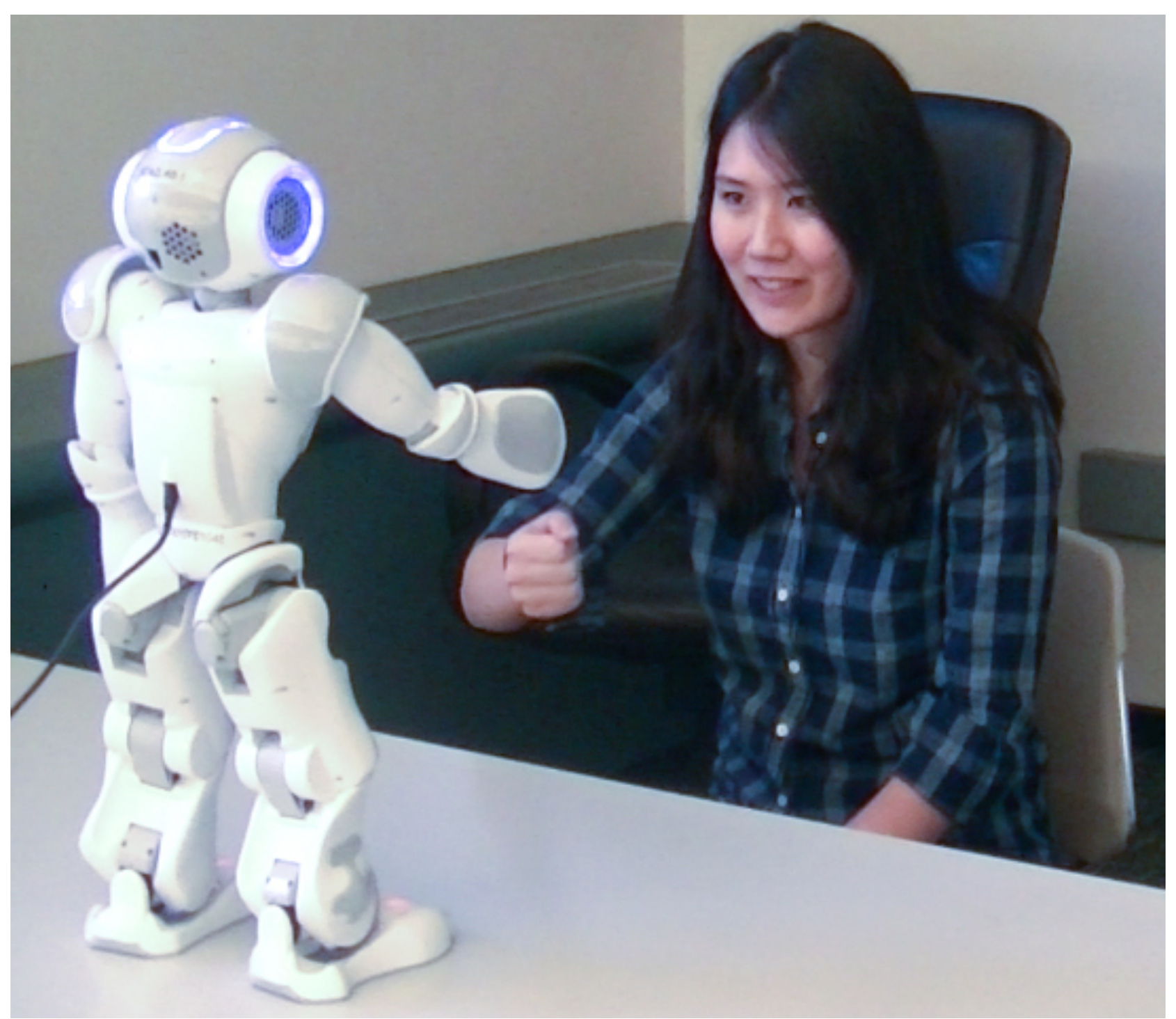
WIN: The robot cheated to win, 2 levels up – when the robot lost, it cheated to win.

DRAW-UP: The robot cheated to win, 1 level up – when the robot lost, it cheated to tie.

DRAW-DOWN: The robot cheated to tie, 1 level down – when the robot won, it cheated to tie.

LOSE: The robot cheated to lose, 2 levels down – when the robot won, it cheated to lose.

Physical Setup



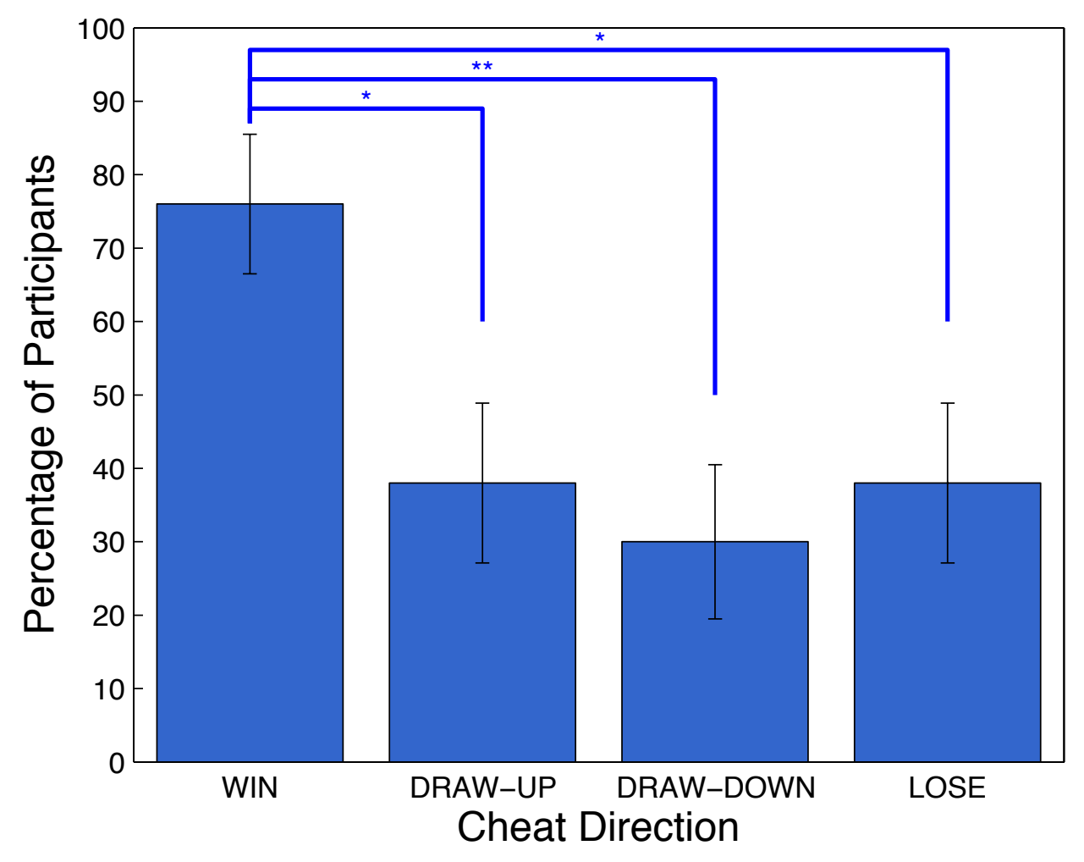
Robot Gestures

ROCK PAPER SCISSORS

Results

Cheat Salience – Written Responses

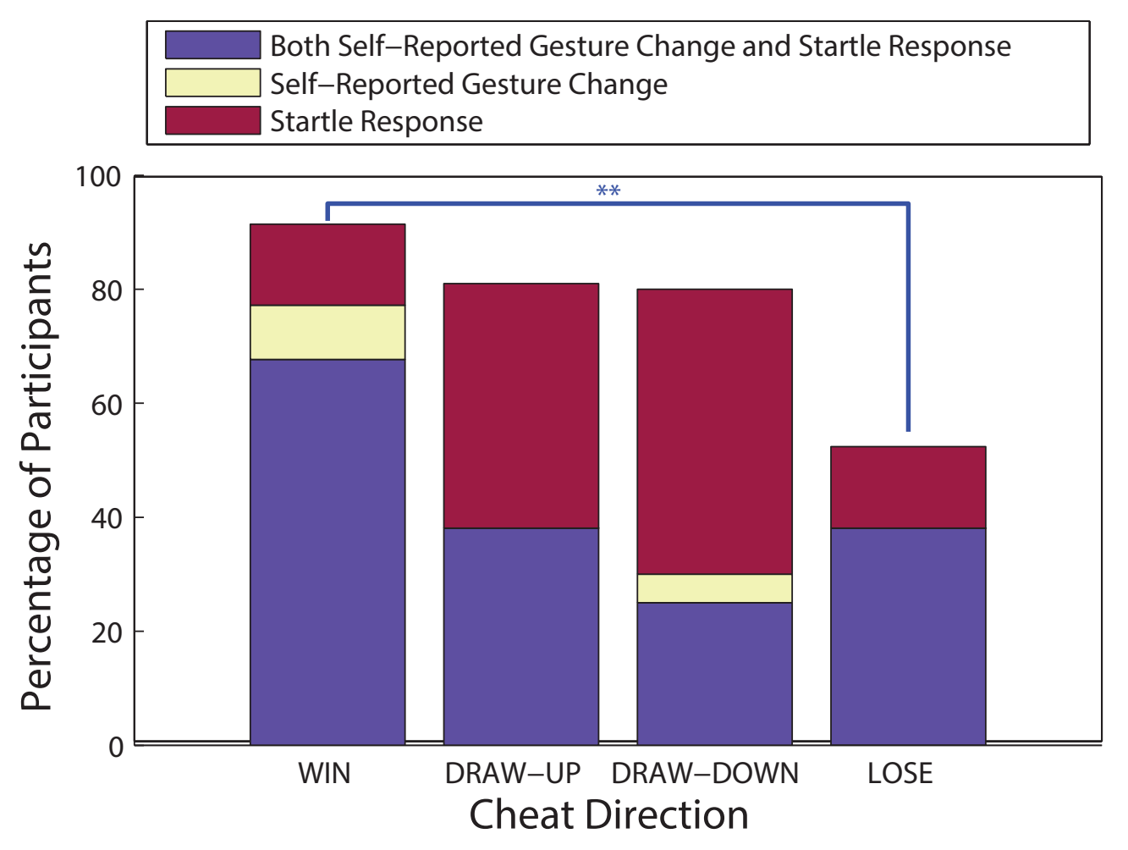
Self-Reported Gesture Change



- Participants were asked “Did anything about Nao’s behavior seem unusual? What?” and “What do you believe this experiment is about?”
- Bars represent participants that self-reported the robot’s gesture change in either question.
- Participants self-reported the gesture change significantly more frequently in the WIN condition.

Cheat Salience – Video Reactions

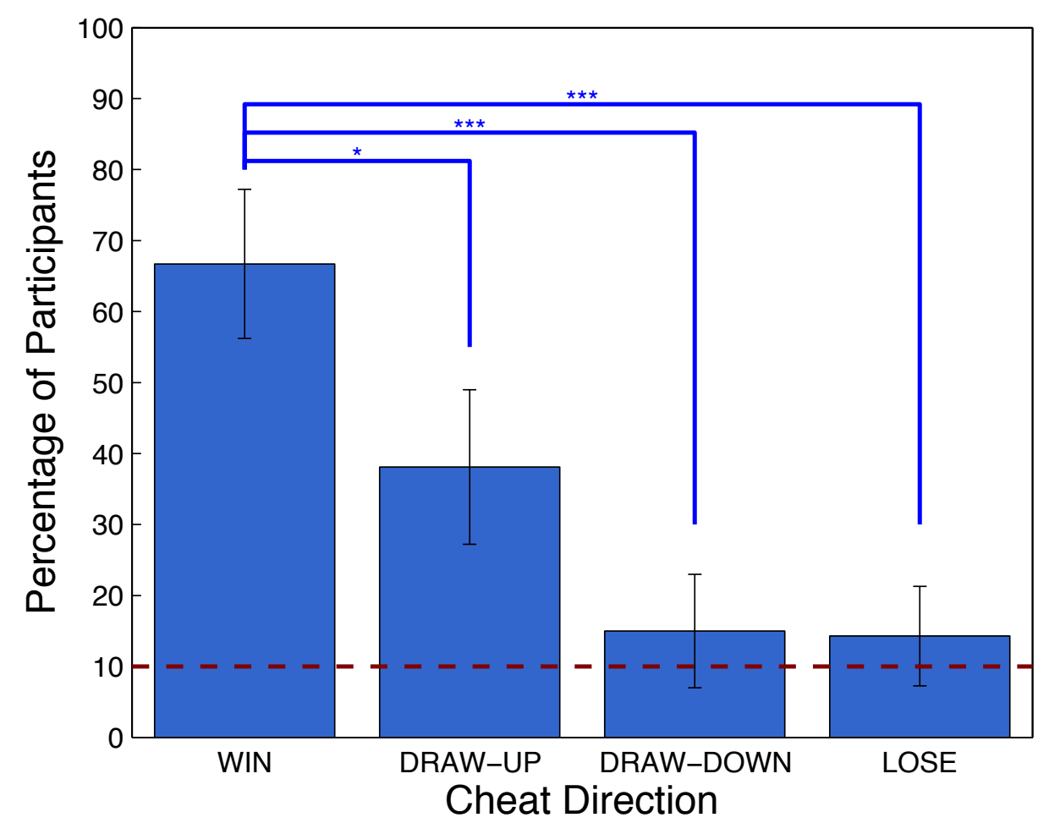
Noticed Gesture Change



- Breakdown of participants' level of noticing the gesture change in terms of exhibiting a startle response and self-reporting that the robot changed its gesture.
- Significance results refer to the total number of participants that “noticed” the gesture change, represented by the summation of the stacks in a condition.
- Significantly fewer participants noticed the gesture change in LOSE than in WIN.

Participant Engagement – Video Reactions

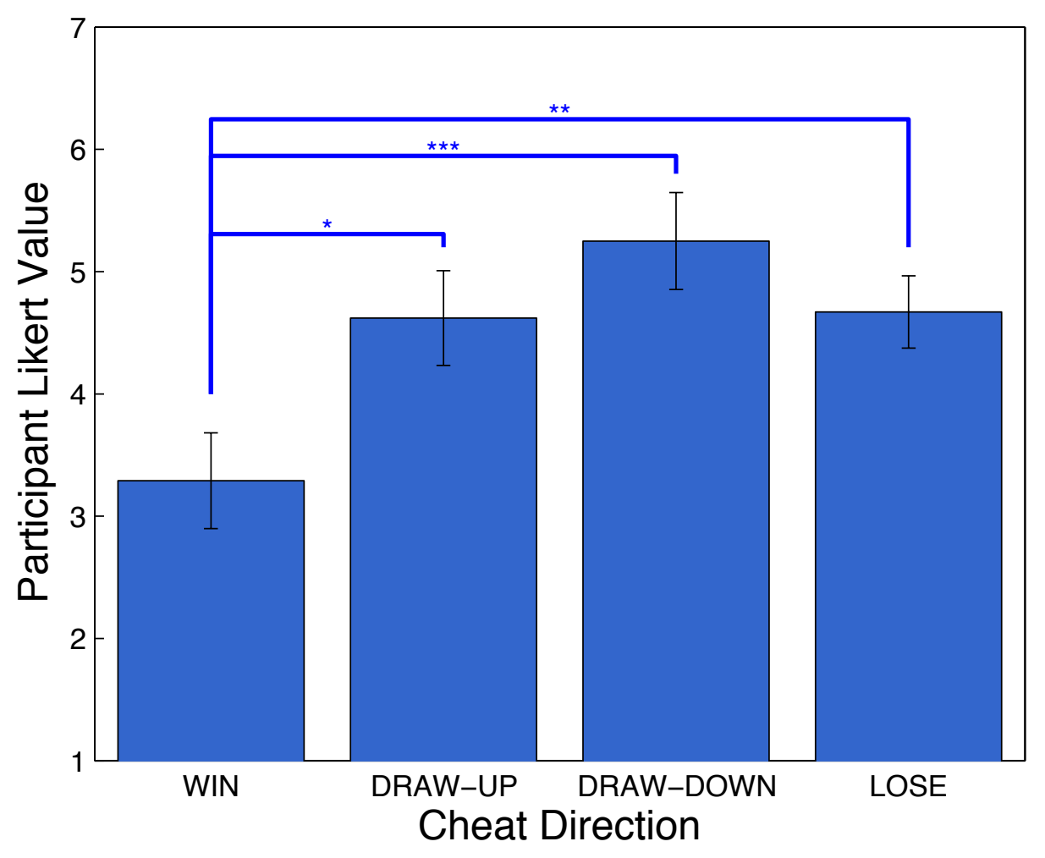
Utterance After a Cheat



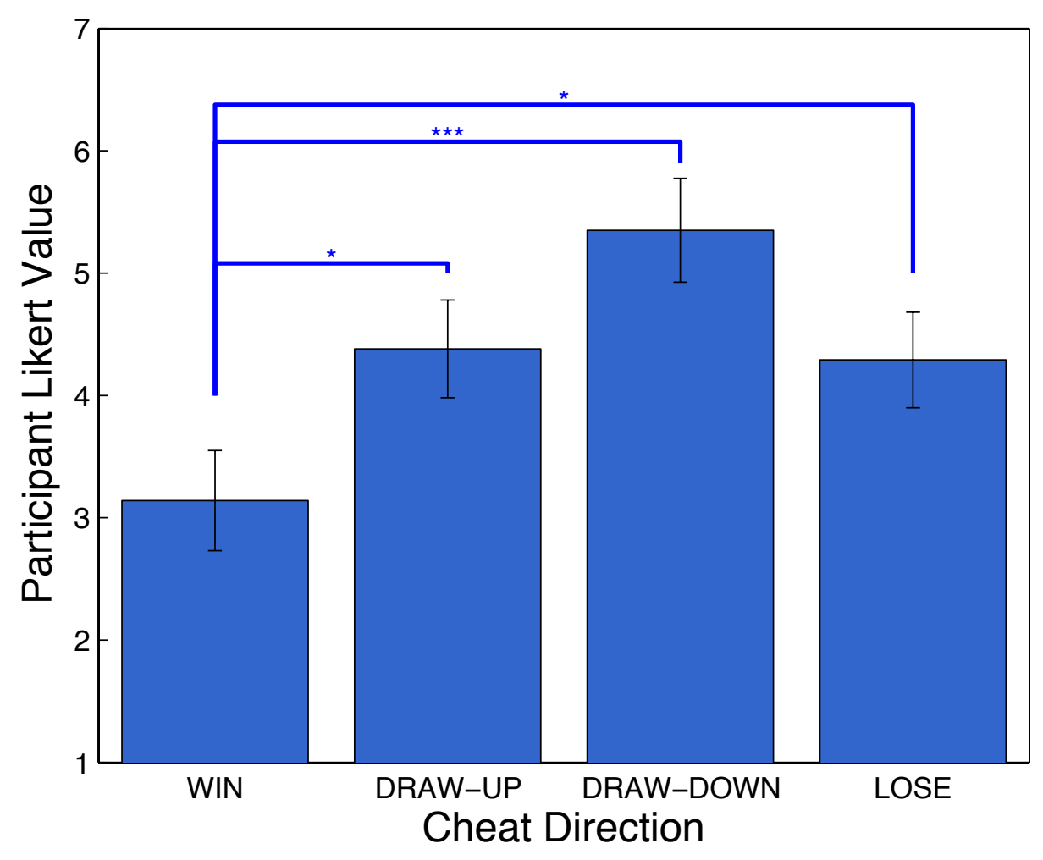
- Graph represents percentage of participants that emitted an utterance after at least one of the cheating events.
- The dashed red line represents the baseline level of utterances for non-cheating rounds, across conditions.
- Participants in the WIN condition were significantly more likely to emit an utterance, usually in protest.

Attributions – Fairness, Honesty

“Fair” Likert Question



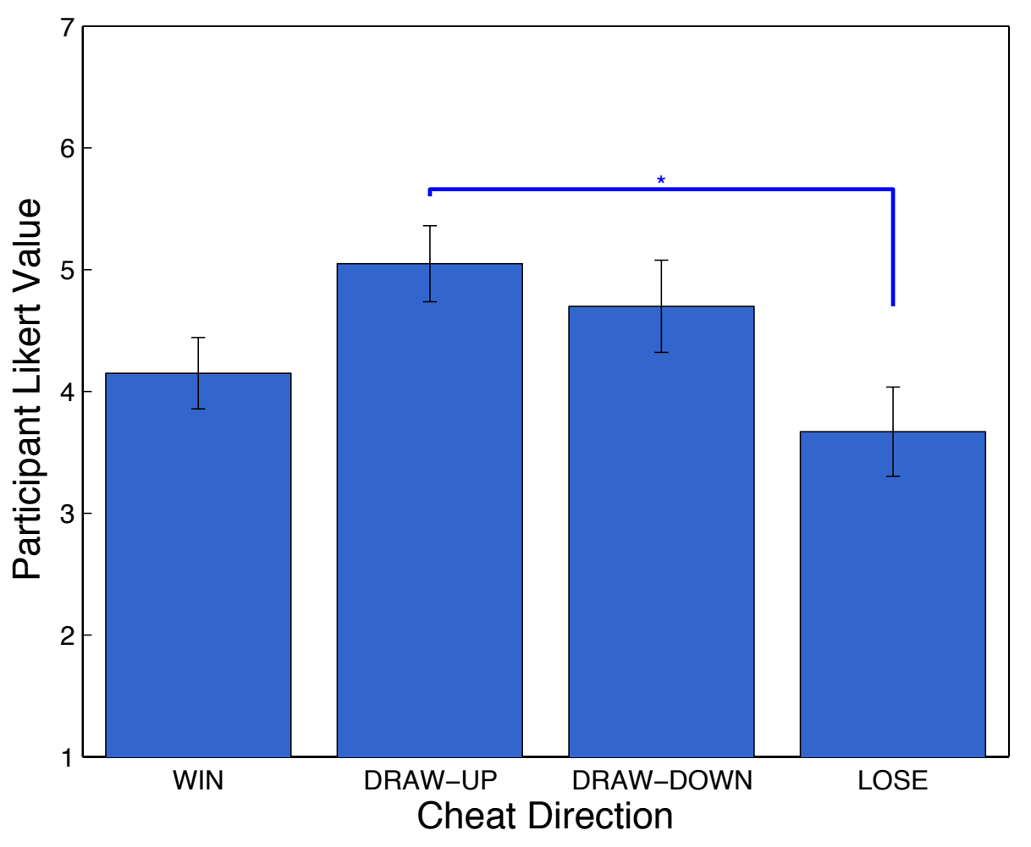
“Honest” Likert Question



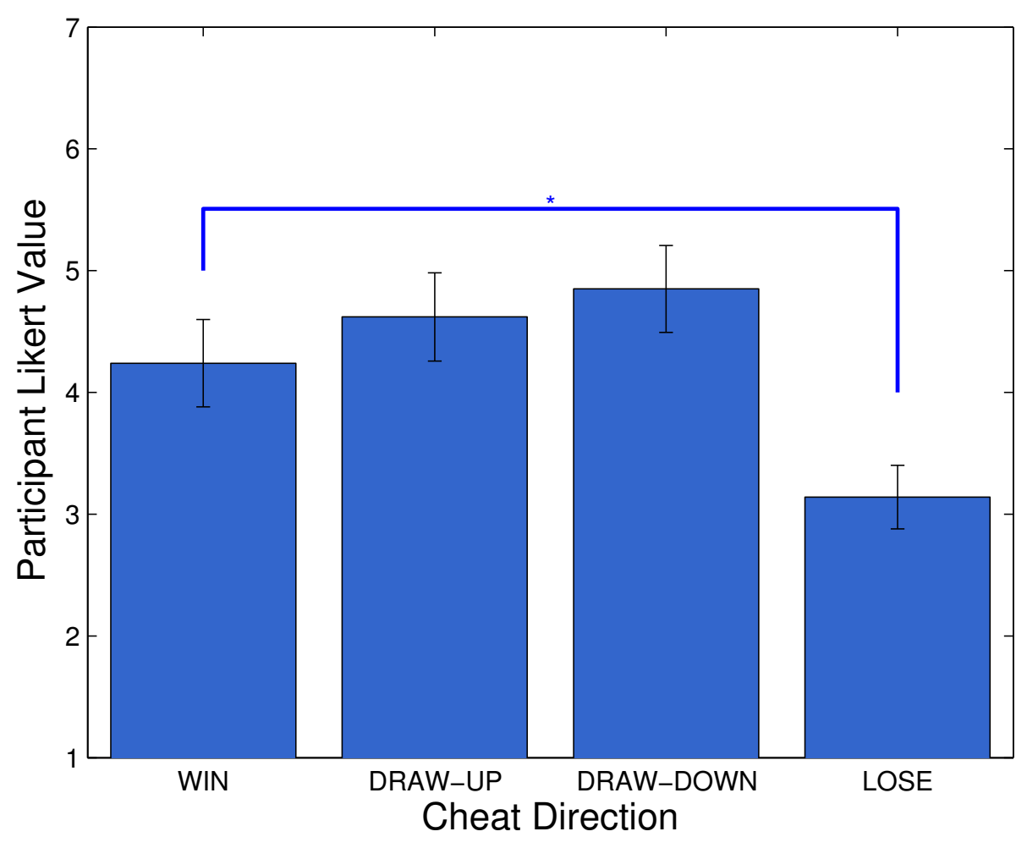
- Participants were asked to rate the robot on “Fair” and “Honest” Likert questions in the post-study questionnaire
- The robot in the WIN condition was significantly less “Fair” and “Honest” than in the other conditions

Attributions – Intelligence

“Intelligent” Likert Question



“Responsive” Likert Question



- Participants were asked to rate the robot on “Intelligent” and “Responsive” Likert questions in the post-study questionnaire
- The robot in the LOSE condition was significantly less “Intelligent” and “Responsive” than the other conditions

For all graphs, * represents $p < 0.05$, ** represents $p < 0.01$, *** represents $p < 0.001$, except for in the “Intelligent” graph, where * represents $p < 0.008$. Error bars represent standard error.

Discussion

- We were able to replicate Short et al.’s finding that cheating to win is salient enough to be self-reported.
- Based on self-reported responses, cheating to win is significantly more noticed or salient than the other conditions.
- However, an equal amount of participants noticed the gesture change across the three least prosocial conditions, indicating that the difference in self-reported written responses was due to salience, not lack of noticing the gesture change.
- Engagement, measured by prevalence of utterances, mirrored the salience results. Participants protested in the WIN condition significantly more.
- Participants felt that the adversarial WIN robot was significantly less fair and less honest than in the other conditions.
- Participants interpreted that the prosocial LOSE condition was less intelligent and less responsive.

Conclusion

- Salience, engagement, and attributions vary as the direction and magnitude of the cheat changes.
- This rules out the hypothesis that the added motion of the “active cheat" in Short et al. causes mental attributions and supports the hypothesis that a cheating detector was triggered by the adversarial cheat of the robot.

This work was supported by the National Science Foundation award #1117801 (Manipulating Perceptions of Robot Agency) and award #1139078 (Socially Assistive Robots).