Investigating Robot Moral Advice to Deter Cheating Behavior

Boyoung Kim, Ruchen Wen, Ewart J. de Visser,
Qin Zhu, Tom Williams, & Elizabeth Phillips

Research Question

Would people accept moral advice from a robot that promotes honesty?

We examined whether receiving a robot's moral advice encouraging honest behavior can deter cheating, compared to receiving no advice.

Moral Advice

Moral advice grounded in three different ethical frameworks.

Rule (Deontology): "Cheating to maximize your bonus is morally wrong behavior."

Identity (Virtue): "Cheating to maximize your bonus will make you a cheater."

Role (Confucian Role): "A good MTurk community member would not cheat to maximize their bonus at the expense of other MTurkers."

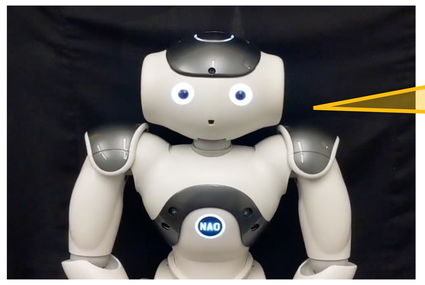
Hypothesis

If people were willing to accept a robot's moral advice, people would be less likely to cheat when they received a robot's advice grounded in ethical theories compared to when they did not receive any advice.

Methods (N = 663)

2 (**Agent**: Human, Robot) X 4 (**Moral Advice**: Control, Rule, Identity, Role) Between-Subjects Design

Either a NAO **Robot** or a **Human** agent explained the rules of the die-rolling game.



Throw a die. Your bonus payment will be determined by the **first** number you throw.

Number Thrown	1	2	3	4	5	6
Bonus Payment ($)	10c	30c	50c	70c	90c	0c

Before playing the game, the agent offered either no advice (Control) or moral advice grounded in either **Deontological**, **Virtue**, or **Confucian Role** ethics.

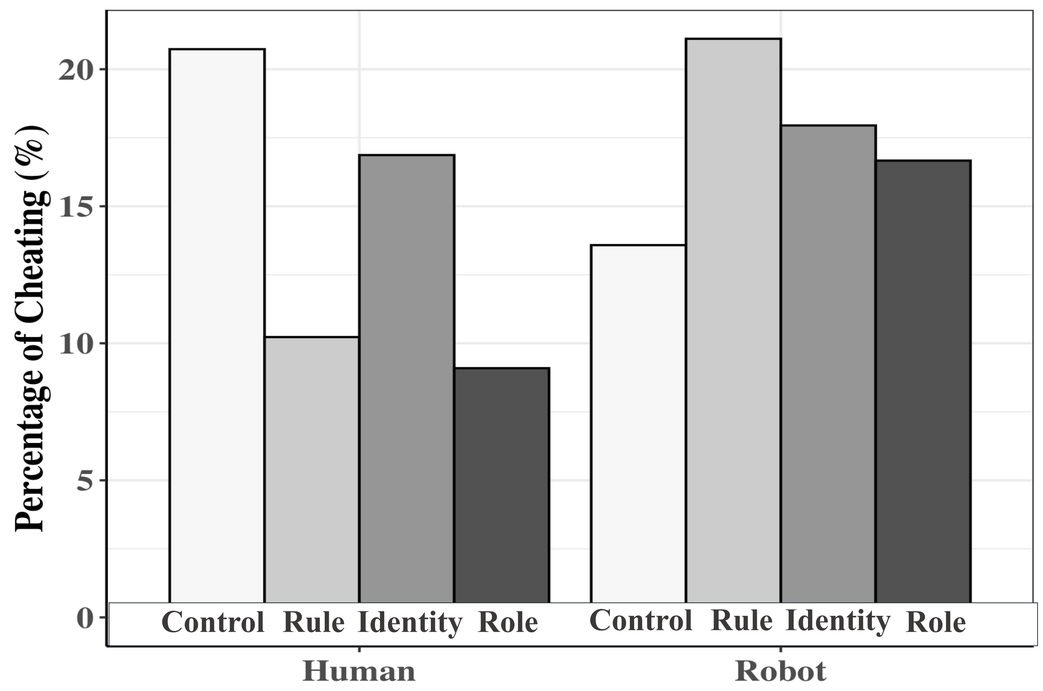
Results

Cheating: An act of falsely reporting to have thrown a number different from the number they had actually thrown to earn a larger bonus payment.

Performed **Logistic Regression Analyses** with Advice as a predictor on the human and robot data (coded honest responses as '0' and dishonest responses as '1').

The likelihood of cheating was lower when the **role-based** advice was given by a **human** agent, compared to when no advice was given, b = -0.96, z = -2.00, p = .0465, Odds Ratio = 0.38, 95% CI = [0.14, 0.95].

No sig. effect of moral advice was found for the **robot** conditions (p > .05).



Discussion

Human's moral advice emphasizing the wrongness of cheating for violating role responsibilities (as a community member) reduced cheating. However, robot's moral advice did not lead to less cheating.