Beyond Programming: Can Robots' Norm-violating Actions Elicit Mental State Attributions?

Joanna Korman
Human-Machine Social Systems Lab
The MITRE Corporation*
Bedford, MA
jkorman@mitre.org

Anthony Harrison
Navy Center for Applied Research in Artificial Intelligence
U.S. Naval Research Laboratory
Washington, DC, US
anthony.harrison@nrl.navy.mil

Malcolm McCurry
Navy Center for Applied Research in Artificial Intelligence
U.S. Naval Research Laboratory
Washington, DC, US
malcolm.mccurry.ctr@nrl.navy.mil

Greg Trafton
Navy Center for Applied Research in Artificial Intelligence
U.S. Naval Research Laboratory
Washington, D.C., US
greg.trafton@nrl.navy.mil

***Abstract*—** Social perceivers often view a human agent's norm-violating behavior as diagnostic of that person's mental states, while behaviors that conform to norms are viewed as less informative. We developed a series of stimulus videos depicting a DRC-HUBO robot engaging in norm-violating and norm-conforming behaviors. We explored the hypothesis that robots' norm-violating actions may invite social perceivers to increase their mental state attributions in a similar manner as they do in humans. Surprisingly, we found that norm-conforming behaviors appear to be at least as conducive as norm-violating behaviors, and perhaps even moreso, to mental state attribution to robotic agents.

Keywords—theory of mind, agency, action explanation, norms, DRC-HUBO

I. INTRODUCTION

Much research on mental state attribution in human-robot interaction focuses on agency cues such as having eyes, a body, and engaging in goal-directed behavior [1]. Considerably less well-studied are the aspects of mental life that are required for people to attribute to robots the same fully agentic action planning that they do to humans, such as *forming a specific intention* or *making a choice among alternatives* (but see [2] and [3] for reference to these topics). Notably, for example, the attribution of a *goal* state ("the robot's goal is to reach the door") may merely refer to the programmer's desired end-state for her robot. It does not necessarily indicate the attribution of flexible reasoning activity – for example, the generation of new goals or wants on the part of the robot – required to form a specific intention to act under novel circumstances.

When social perceivers consider the mental activity behind a human's decision to act, they often consider whether or not the behavior in question conforms to a norm. Specifically, norm-violating behaviors are often seen as diagnostic of underlying mental activity – a specific, conscious choice a person made to act in a particular situation – while norm-conforming behaviors are considered to be less diagnostic, sometimes involving merely habitual behavior that reflects more about the pervasiveness of the norm itself than about the mental states underlying any individual's decision [4, 5]. For example, when people observe a person throwing an empty container into a trashcan, they may be more likely to explain such behavior by appeal to the norm itself (e.g., *he threw it in the trash because that's where trash is supposed to go*) – than by generating any mental state that is unique to the particular agent or behavior (*he threw the container in the trash because he wanted to show off to his girlfriend what a good citizen he was*). When a robot performs a norm-conforming behavior like throwing away trash, this behavior may also be easily understood by the social perceiver as fulfilling a goal that is either totally pre-programmed into the robot, or is at least predictably activated in response to an environmental cue.

The present study provides a first examination of how people interpret norm-violating behaviors from the observed, physically instantiated actions of a robotic agent, and whether such violations are considered diagnostic of situation-specific mental state reasoning. If robotic agents are, like humans, subject to attributions of these more complex and specific intentions, norm-violating behaviors performed by a robot should reveal these attributions. The present study explored this phenomenon by creating a set of stimulus behavior videos that depict a robot engaging in norm-violating behaviors in everyday social contexts.

II. METHODS

To examine the role of norm violations in eliciting judgments of rich mentalistic activity, we created videos of a DRC-HUBO robot on wheels in three distinct scenarios: throwing away a piece of trash, getting into a line of people, and entering an elevator where a person was already standing. Each of the three scenario videos was presented in three distinct versions: norm-violating (experimental) and non-norm-violating (control), as well as a third ("mistake") condition, which served as an additional control, in which a norm violation was present but appeared unintentional.

For example, in the scenario depicting the line, participants viewed a registration line that went around a corner (See Fig. 1, frame 1). The non-norm-violating condition then portrayed the robot simply wheeling from

*The author's affiliation with The MITRE Corporation is provided for identification purposes only and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author. Public Release: *18-4603*. Distribution Unlimited.

offscreen to stand behind several people in a spot at the end of the line (2a). In contrast, the intentional norm violation condition depicted the robot cutting in line: the robot wheeled itself past several people on the line and straight in front of the person at the head of the line (2b). Finally, the mistake condition (2c) depicted an unintentional violation: the robot wheeled to stand behind one person at the front of the line, but still came to stand in front of a group of people who were occluded around a corner.

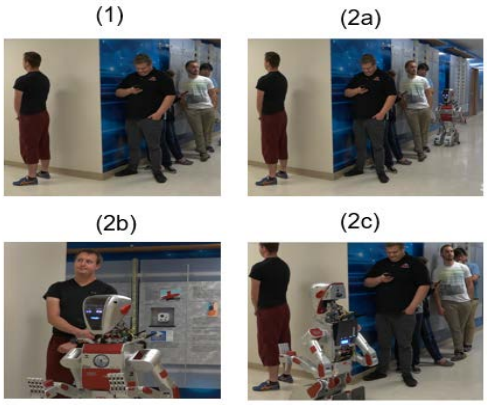

Fig. 1

At the conclusion of each video, participants were invited to explain why the robot had performed this behavior, and to evaluate (on a 0 to 6 scale) whether the robot completed the behavior *intentionally*, whether the robot *chose* to perform the behavior, whether the robot *wanted* to perform the behavior, whether the robot was *aware* of performing the behavior, and whether the robot could have chosen *not* to perform the behavior. Each participant viewed one of the nine videos and answered the above questions. 82 participants completed the study on Amazon Mechanical Turk.[1]

III. Results

A. Mental state ratings

We first performed five separate ANOVAs comparing means on the rating scales of interest (want, awareness, intentionality, choice, and ability to have chosen otherwise). We treated condition (violation, non-violation, mistake) as a single factor. Although a significant difference among conditions emerged for the *want*, *awareness*, and *intentionality* questions, $Fs(2, 79) > 5.89$, $ps < .01$, surprisingly, this difference was not in the predicted direction: the greatest attribution of mental states for all three of these measures took place in the *absence* of any norm violation (e.g., $M_{\text{Non-violation}} = 5.32$), and the non-violation condition differed significantly from the intentional violation and mistake conditions (Tukey's HSD tests for all three measures, $ps < .05$), which were statistically indistinguishable from one another ($M_{\text{Int. violation}} = 3.52$; $M_{\text{Mistake}} = 3.17$). The the two choice questions showed no significant variation across conditions, $Fs (2, 79) < 1.65$, $ps > .19$.

B. Qualitative measures

Responses to the open-ended explanation prompt ("Why did the robot [perform the behavior in question]?" were classified into one of several categories, including those that (1) clearly cited mental states preceding a planned action (i.e., "the robot wanted to enter the elevator"), and (2) alluded to the robots' previous programming by a human. In the present analysis, explanations that cited an error in the robot's perception (possibly mentalistic, but unrelated to the agent's action plan) or other background factors (also non-mentalistic) are grouped together in a third category. A chi-square test for independence revealed no significant relationship between response category and condition, $\chi^2(4) = 1.80$, *ns,* and the number of programming vs. mental state responses were comparable across all three conditions.

IV. Discussion

On a range of measures of the mental state activity important for action planning – awareness, desire, and intention, as well as on open-ended explanations of the robots' behaviors – we found that, in contrast to their attributions for humans, people attribute the same number, or even fewer, mental states to robotic agents that engage in norm violating actions. What can account for this discrepancy?

One possibility is that our participants found the robotic agents' norm violations – which they watched unfold on video – to be implausible. For example, additional free response data revealed that most participants who saw the intentional "trash" norm violation – the robot holds a piece of trash over the trashcan, stops for a moment, and then moves it away from the can and drops it on the floor – did not recognize this as a norm violation ("littering") at all. Instead, they regarded it as an unintentional violation ("accidentally dropping trash on the floor"). It may be that because people tend to view robots in terms of low-level, programmable goals that serve clear human ends, they may be disinclined to interpret such scenarios as instances of intentional norm violations at all. Social perceivers may view robots as relatively insensitive to social norms, perhaps because they lack the sorts of goals and desires that would ever incline them to break such norms in the first place.

[1] Abubshait, A., & Wiese, E. (2017). You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human-robot interaction. *Frontiers in Psychology*, 8: 1393.

[2] Ullman, D., Leite, I., Phillips, J., Kim-Cohen, J., & Scassellati, B. (2014). Smart human, smarter robot: How cheating affects perceptions of social agency**.** In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society.* Quebec City, Canada. July 23-26, 2014.

[3] Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*, 14, 1962.

[4] Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116,* 1, 87-100.

[5] Monroe, A., Dillon, K., Guglielmo, S., & Baumeister, R. (2018). It's not what you do, but what everyone else does: On the role of descriptive norms and subjectivism in moral judgment. *Journal of Experimental Social Psychology,* 77.

[1] Supplementary materials can be viewed at: https://osf.io/e85fw/