"Fight-or-Flight": Leveraging Instinctive Human Defensive Behaviors for Safe Human-Robot Interaction

Karthik Mahadevan
University of Calgary
karthik.mahadevan@ucalgary.ca

Sowmya Somanath
OCAD University
ssomanath@faculty.ocadu.ca

Ehud Sharlin
University of Calgary
ehud@ucalgary.ca

ABSTRACT

Maintaining the safety of humans is of paramount concern in the field of human-robot interaction. We employed a Research through Design (RtD) approach to explore better HRI safety mechanisms. We conducted a preliminary design study where we presented a group of designers various scenarios of different robotic platforms acting unsafely. Our findings indicate that participants mapped human responses to unsafe robotic interfaces, to natural human defensive behaviors in response to varying levels of threat stimuli. Based on preliminary findings, we suggest leveraging the instinctive human ability to react to dangerous situations as a fail-safe mechanism to the robot's own built-in safety methods.

KEYWORDS

Safety in Human-Robot Interaction

ACM Reference Format:
Karthik Mahadevan, Sowmya Somanath, and Ehud Sharlin. 2018. "Fight-or-Flight": Leveraging Instinctive Human Defensive Behaviors for Safe Human-Robot Interaction. In *HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, March 5–8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3173386.3177004

1 INTRODUCTION

Ensuring the safety of humans is a crucial concern in human-robot interaction (HRI) research. In their extensive survey of HRI safety methods, Lasota et al. [4] identify that both physical safety and psychological safety have to be satisfied throughout the interaction to ensure overall safety. They define physical safety in HRI being met when there is no unintentional or unwanted contact between the robot and the human. Currently, safe HRI research is actively engaged in improving the technical competencies of robots so they can safely interact with humans.

As robots become increasingly autonomous and independent entities, humans will have less of a role to play in intervening [5]. While the intelligence powering such robots is rapidly improving, it is far from matching human intelligence when faced with unpredictability. An autonomous robot's functionality can be limited by the quality of data it senses, algorithms governing its many behaviors, and its computational prowess. Limiting our work to physical safety, we explore human intervention as an additional fail-safe mechanism for scenarios of robot errors which can place people in a potentially dangerous situation.

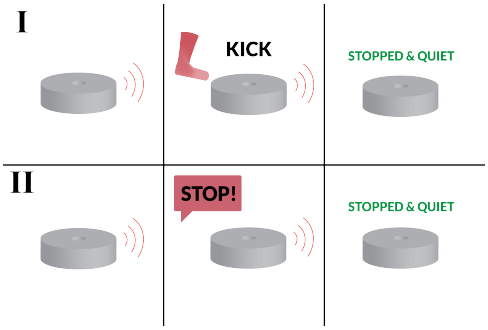

Figure 1: Participants' example low-fidelity design ideas reflecting on human response to a robotic vacuum cleaner (Roomba) acting unsafely.

The simplest form of intervention is an industrial "kill switch", which is designed to power down equipment and machinery immediately. While kill switches can completely disable a robot and avert possible danger, we want to explore other techniques that will allow the robot to quickly correct its unsafe behavior and continue performing its tasks in a safe manner (unlike kill switches).

When humans are faced with a seemingly dangerous situation, they respond by displaying specific defensive behaviors, such as "fight-or-flight" [1]. We hypothesize that by incorporating the natural instinctive defensive behavior of humans to threat stimuli, we can maintain safe human-robot interaction in the event that the robot's own safety methods fail to act.

2 RELATED WORK

In [4], Lasota et al. survey four major methods to maintain safe HRI: (i) safety through control, (ii) safety through motion planning, (iii) safety through prediction, and (iv) safety through consideration of psychological factors. Most of these techniques aim to improve a robot's technical competencies to enhance safety during interactions with humans. However, a robot may still behave in an imperfect manner from time to time. For example, robot safety methods that require perception could suffer from a momentary lapse in obtaining sensor data or an inaccuracy in collecting it, preventing a robot from activating the safety mechanism in a timely manner.

Work in neuroscience [1] has found that several factors influence the human response to perceived threat, including but not limited to escapability, distance from a threat, and ambiguity of the threat stimuli. Overall, varying levels of threat stimuli are shown to provoke different types of human defensive behaviors (such as "run away", "attack", and "yell, scream, or call for help").

Incorporating the innate human response as a fail-safe mechanism for safe HRI is less explored in the literature, and is the focus of our work.

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03.
https://doi.org/10.1145/3173386.3177004

3 STUDY DESIGN AND FINDINGS

We explored the innate human response as a fail-safe mechanism in safe HRI using a Research through Design (RtD) approach [6]. We asked several designers to engage in a design activity that required them to reflect through design and prototyping, on dealing with various unsafe robot behaviors across several scenarios. We recruited five participants experienced in interaction design, in the age range of 18-35. All participants were recruited on a university campus through word of mouth, and received a cash remuneration of $20 at the end of the study. Participants were seated to mimic the arrangement of a round table discussion. We briefed participants on potentially dangerous scenarios involving four autonomous robotic platforms: (i) a Baxter, a large humanoid (Rethink Robotics), (ii) a self-driving car, (iii) a Roomba robotic vacuum cleaner, and (iv) a NAO humanoid. We chose these disparate platforms to center discussions around differing size, form, functionality, and perceived intelligence. We introduced each platform to participants through a brief description and a short video showcasing its use cases and functionalities. For each platform, we described a few scenarios where the robot displays unsafe behavior. We also encouraged participants to think of scenarios beyond these examples. We first asked participants to reflect on how they would respond to a situation where the robot could physically harm them. Then, we asked participants to design an interface or an interaction technique to communicate their concern to the robot, along with the robot's response to the concern. We asked participants to demonstrate their new interaction techniques in the form of low-fidelity prototypes, sketches, or enactments. To help participants with this process, we provided them with office stationary such as pens, sticky notes, and sheets of paper. Figure 1 presents a sample of a participant-provided solution to unsafe behavior by a Roomba robot.

To analyze the data, we employed qualitative research methods by assessing the video-recorded study session and applying open coding to identify common threads in participants' discussion [2, 3]. Some examples of the codes we used were, "initiating physical contact with robot", "form of robot", and "rate of approach".

Participants identified two major criteria affecting the threat they ascribed to a potentially dangerous situation, namely, the form of a robot, and the rate of its approach. Participants strongly associated form with the potential danger posed by a robot. For example, a small and harmless Roomba could drive over a person's leg but wouldn't typically inflict a significant amount of discomfort or pain on them. In contrast, participants obviously attributed significantly more danger to the self-driving car and the Baxter, both of which can potentially be perceived as menacing due to their physical size or appearance. Participants considered the robot's rate of approach as another significant factor when perceiving its level of threat. Participants felt that they would be more comfortable in a potentially dangerous situation posed by a Roomba because of its slow and predictable approach. On the other end of the spectrum, participants cited the self-driving car's ability to accelerate and achieve high speeds as reasons for being more fearful of any physical contact with it.

Participants proposed initiating physical contact with the Roomba as a possible solution to instantaneously send it a stop command in response to a potentially dangerous situation. Participants also did not expect to feel any remorse when physically assaulting the robot in this context: "I don't feel bad about kicking it if it is endangering me" [P1]. From the human perspective, this idea encourages actively seeking physical contact to send a message to the robot, running counter to the robot's goal of avoiding unwanted contact. Similar ideas emerged when we probed participants about the NAO and the Baxter. Conversely, participants vehemently opposed the idea of sending a message to the self-driving car through physical contact, probably due to both its size (giving it the ability to seriously injure a human) and rate of approach. When we questioned participants on their response to a potentially dangerous situation with a self-driving car when crossing a street, participants opted either not to cross (and wait until the vehicle passed) or to scurry across to the other side of the crosswalk if they had already begun crossing. To communicate concern to such a platform, participants suggested indirect methods such as gesturing or altering their body pose. Participants also suggested the use of voice to communicate fear in a potentially dangerous situation. As an example, participants wanted the Baxter to be able to respond to screams and phrases that could be construed as panic.

4 DISCUSSION AND FUTURE WORK

Our findings suggest interesting parallels between the work on human defensive behaviors to threat stimuli [1] and the ideas participants suggested to incorporate the human response. For example, kicking the Roomba or tapping the NAO are akin to attack responses, fleeing a crosswalk in the presence of self-driving cars mirrors flight, and screaming when on the verge of being hit by a Baxter resembles expressing fear through voice. Although a Roomba would probably never pose the same level of physical danger as that of a Baxter, it could still elicit an instinctive response from a human. Our findings hint that just as humans display defensive behaviors when threatened by other humans, and non-human mammals display antipredator behaviors, humans may display similar behaviors in the presence of robotic entities.

Given that the human response is informed by natural selection and is well developed, we think robots should be designed to recognize such behaviors and respond to them as an additional safety measure when their own safety mechanisms fail. Since this work is preliminary, we would first need to verify that these behaviors actually manifest in an experimental study setting. While nontrivial, the focus will then be to develop a framework on a suitable robotic platform to allow it to sense and respond to specific human defensive behaviors. We can then study and contrast the effectiveness of such an approach with more orthodox safety methods.

REFERENCES

[1] D Caroline Blanchard, April L Hynd, Karl A Minke, Tiffanie Minemoto, and Robert J Blanchard. 2001. Human defensive behaviors to threat scenarios show parallels to fear-and anxiety-related defense patterns of non-human mammals. *Neuroscience & Biobehavioral Reviews* 25, 7 (2001), 761–770.

[2] Kathy Charmez. 2006. Constructing grounded theory. (2006).

[3] Adam Fouse, Nadir Weibel, Edwin Hutchins, and James D Hollan. 2011. ChronoViz: a system for supporting navigation of time-coded data. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 299–304.

[4] Przemyslaw A Lasota, Terrence Fong, Julie A Shah, et al. 2017. A Survey of Methods for Safe Human-Robot Interaction. *Foundations and Trends® in Robotics* 5, 4 (2017), 261–349.

[5] Holly A Yanco and Jill Drury. 2004. Classifying human-robot interaction: an updated taxonomy. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, Vol. 3. IEEE, 2841–2846.

[6] John Zimmerman, Erik Stolterman, and Jodi Forlizzi. 2010. An analysis and critique of Research through Design: towards a formalization of a research approach. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*. ACM, 310–319.