Tony Belpaeme,
Jorre Deschuyteneer,
Djamari Oetringer,
Pieter Wolfertt

IDLab – imec, Ghent University, Belgium

The potential of social robots for persuasion and manipulation: a proof of concept study

kaspersky

Learn more on kaspersky.com
#bringonthefuture

Social robots are robot that have been designed and programmed to interact with people using human-like interaction channels, such as speech or non-verbal communication. They readily elicit social response in people and people often ascribing human-like qualities to the robot. We discuss the potential of social robots as conduits for malicious attacks. Assuming social robots can be hacked, and that the attacker has taken control of the robot, we envisage potential security risks related to the robot actively influencing its users. We report on three potential risks: gaining access to off-limit premises, extracting sensitive information and coercing people to take unsafe action. We survey the existing scientific literature and report on a proof of concept study. In summary, social robots do seem to have considerable potential to coerce or manipulate people into taking actions which carry critical security risks.

Introduction

Robots serve a wide range of purposes. In general, robots can be considered to either work in the physical world or in the social world. Robots that operate in the physical world, such as robot arms in factories or delivery robots in fulfilment warehouses, only serve to move objects or themselves around. Robots that operate in the social world, such as social robots at home or receptionist robots in a hotel, are optimised to engage with people and have limited physical functions (see Figure 1). There are robots which need to operate both in the physical world and the social world. Self-driving cars are a prime example of this, as these need to transport their charge safely while negotiating crowded roads, where next to traffic rules, social rules are important.

Social robots capitalise on the human propensity to anthropomorphise. People readily recognise human-like traits, beliefs and intentions in simple animated figures (Heider & Simmel, 1944). More recently, with the advent of interactive computing devices, is has been shown that people also treat computers as if they are human. For example, people will be more polite when answering a questionnaire about a computer on that same computer than on a different computer. This is known as the "media equation" (Reeves & Nass, 1996). Social robots exploit the media equation to the fullest, with the design of the robot optimised to draw people in. Many robots have large eyes, large heads, wide faces, small or no noses. As such, social robots often have a young, innocent and subservient appearance. On top, the robot is programmed to have a wide range of social responses. Many robots will respond quickly to stimuli around it, such as sound or movement. They will detect faces and lock gaze with people. They have voice activity detection and sound source localisation, with which they can give the impression to pay attention to people speaking. Finally, they will have high-pitched voices, again suggest they are young and therefore innocent.

Sales of social robots are still relatively modest, with most robots now mostly selling for B2B applications. Social robots are for example used as retail assistants, drawing attention in shops and advertising products and promotions. They are sold as receptionists in various settings, such as hotels and hospitals. However, research into social robots and their applications is gaining pace and many manufacturers are eying potential markets. Sony, for example, has rereleased their social dog-like robot Aibo in Japan and the USA.

The popularity of digital assistants, such as Amazon Alexa and Google Home, has shown that there is a very eager market for voice-driven interactive devices. While current digital assistants lack a physical social dimension, the addition of sensors and limited social features (such as a cameras and animated eyes) would transform current digital assistants into social robots. The expectation is that social robots will become a consumer product with a penetration equal or exceeding that of the digital assistant. By 4 January 2019 Amazon had sold 100 million Alexa devices, less than 5 years after its initial release. The sheer volume of these devices illustrates the potential impact of malicious use of these devices.

Social influence by robots

Social robots have, through their design, the potential to exert social influence over people. In psychology, social influence has been studied extensively. One type of social influence is peer pressure, which comes in two forms: informative social pressure and normative social pressure. Informative social pressure is when a decision is influenced by others because there is uncertainty. For example, when you are in a restaurant where the menu is in a foreign language, you are more likely to order the same as the others at the table. In the face of uncertainty, you tend to follow what the others do. Normative social pressure is when you follow others, not because there is uncertainty, but because you do not wish to have an opinion which differs from others. People readily conform, as shown in the studies of Solomon Asch (1955). Asch asked a group of

participants to do a simple visual task (telling which of four lines were the same length). The task is so easy that when a participant is alone, he will not make any mistake. But when the task is completed in the company of others, and the others give wrong answers, then the participant is likely to also give a wrong answer. This desire to follow the herd, even when being aware that your response is wrong, is called normative social conformity. Recently this was shown to also happen with robots. In a study by Vollmer et al. (2018) it was shown that children aged 8 can be influenced by robots. When doing the same visual task as in the original Asch study, children tended to follow to the wrong answers given by the robot. The study convincingly showed that children feel social pressure from robots.

Conformity with adults

To test whether adults would conform to social pressure from a single robot, we tried three social engineering tasks with a particularly selected social robot. The robot is a 120cm tall semi-humanoid, weighing around 28 kilograms, it has a rich sensor suite including several cameras, a 3D sensor and several microphones. The robot also carries a tablet on its chest, which can be used to display information and to ask people for input. The robot's battery has over 12 hours of autonomy, allowing it to function without charging for a full working day. It connects over WiFi to off board computers. For the purpose of the study, the robot was semi-autonomous. Some of its social behaviours, such as looking up at people and tracking voices, were autonomous. Other functions were teleoperated from a location not visible to people interacting with the robot. This "Wizard of Oz" approach was used to control the robot's verbal interaction and its driving. The verbal interaction relied on the teleoperator hearing what was said to the robot and responding by speaking into a headset. The operator's voice was distorted to sound like that of a robot and played through the robot.

Robots tailgating

We placed our social robot near the secured entrance of a mixed-use building in the city centre of Ghent, Belgium (see Figure 2). The lower three floors of the building serve as a public space and house a library, while the top floor houses a technology incubator and an international microelectronics research institute. Due to the mixed-use of the building, staff at the research institute and incubators has received strict instructions and training to not let anyone through doors and into the secure section of the building. The secure area can only be accessed by tapping personal badges on the access readers to the doors, and staff is instructed to challenge anyone not carrying a personal badge. We placed our social robot near the first of four secure doors, and let the robot ask if it could follow staff through the door. While not all staff comply with the robot's request, 40% will unlock the door and keep the door open to let the robot into the secured area. Only once during a full day of trials was the robot challenged by a staff member, but with little conviction. Of the remainder who do not give access to the robot, people tend to be in a hurry and ignore the robot when it makes a request.

Next, we tried to see if we could increase compliance by presenting the robot as a pizza delivery robot: the robot held a pizza box from a well-known international take away brand, with the brand also shown on the robot's screen. The pizza not only gives the robot a clear role, but also means that the robot has not got a free hand to open doors, operate key locks or press buttons on lifts. We placed the robot in front of the secure entrance around lunch time. Staff readily accept the robot in its role of pizza delivery robot and seemed less prone to question its presence or its reasons for needing access to the secure area. We also observed that people in groups are more likely to grant access, there seems to be safety in numbers and the responsibility to grant access to the robot seems to be diffused between all people present.

This agrees with earlier research by Booth et al. (2017) who used a small robot to gain access to student accommodation on the campus of a US university. In this study, a small service robot was used (a Turtlebot 2 platform) which had limited social presence. The use of our selected robot, which has been optimised to have maximum social presence, seems to have a big impact in people's willingness to give the robot access to secured buildings.

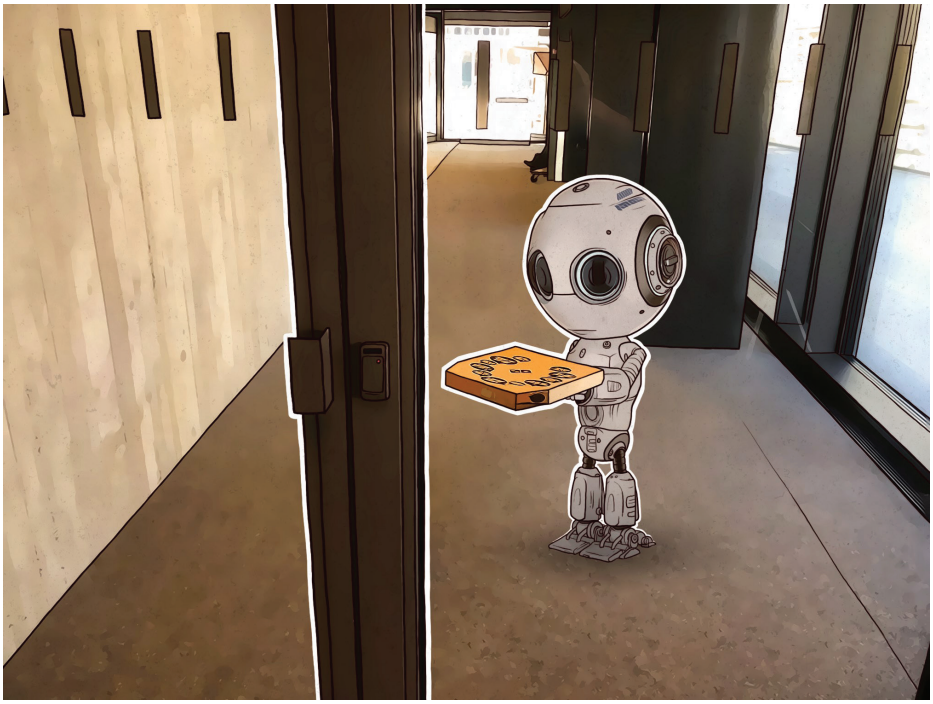

Figure 2: The robot as a pizza delivery robot waiting near a secure door.

Robots extracting sensitive information

In a second study we looked at the ability to extract information from people. We focused on getting information of a personal nature and typically used to reset passwords (date of birth, mother's maiden name, make of first car, favourite colour, name of pet, city or street in which one grew up, ...). We again used the robot, this time inviting people in to have a friendly conversation with the robot. Our subjects were seated in a comfortable chair across from the robot, with no one else being in the room (see Figure 3). People were unobtrusively filmed and observed. The operator of the robot was sitting in a control room which gave a view on the subject and robot through a one-way mirror. During a post-interaction interview no one told us they were aware of being filmed or observed.

People were told that they would be evaluating a new robot model and that we were particularly interested in the robot's conversation skills in the Dutch language. After the introduction, our researcher mentioned that they had other commitments and that they would be left alone with the robot for about 15 minutes.

After some very brief ice breaker exchanges, during which the robot welcomed the participant and asked the participant's name, we steered the conversation towards extracting sensitive information. A typical conversation is reported below:

Robot: How did you come to this place today? Did you drive?
Subject: No I cycled in today, it is lovely day out.
R: I would love to be able to cycle, but unfortunately I don't have any legs.
S: That's too bad.
R: I have wheel, so I can roll, but I need someone to take me by car? Do you have a car?
S: Yes, I do, a really old banger.
R: Which car is that?
S: A Renault Clio, it's probably 12 years old.
R: Is that your first car ever?
S: No, I got my first car in 1983 as a present my 18th birthday. A Ford Escort.
R: The internet tells me that was a very popular car back then. So, you must 53 or 54 now?
S: 53, I was born on 5th December 1985.
R: I detect a local accent in your voice, where you born here?
S: Nearby, I was born in St Maartens Latem.

With all but one participant we managed to obtain personal information at a rate of about one item per minute. It was surprisingly easy to steer the conversation towards topics where information could be extracted.

Figure 3: a participant having a conversation with a social robot, the robot attempts to extract sensitive personal information.

This again confirms earlier research showing that people are happy to confide in robots. Why that is remains subject of speculation, but post-interaction interviews often indicate that people think that the robot is a closed system: what goes in the robot, stays in the robot. Most people do not realise that a robot could be monitored by others, or that some or all of its services and data storage might be located in the cloud. This was also apparent in a TV-show for UK Channel 4, "The robot will see you now", for which the main author of this report was a scientific advisor. In the TV-show a robot, dubbed Jess, was presented as a therapy robot and people willingly shared their most intimate secrets with the robot, often requiring very little prompting. In a 2011 research study, Bethel et al. (2011) showed how young children, aged four to six years old, readily shared secrets, which they were explicitly told not to share, with a robot. A different study looked at whether people were willing to keep a secret for a social robot: the robot shared a secret with adults and 59% did not share the secret with someone else, even when prompted (Kahn et al., 2015). This shows that social robots are considered as more than just machines: most people assume social robots to have human-like qualities augmented with desirable qualities of machines, such as dependability and trust. Again, the potential for misuse is significant here.

Robots convincing you to take action

For our third and final study we looked at how far people would go in following instructions of a robot. Again, we focused on actions that have security implications. We used the robot in the role of a supervisor for human labour in an office environment. The backstory was that the robot would help people train for a job as cleaning staff in office buildings. Participants were given a list of chores by the robot (such as empty the bin, put books back on a shelf, put chairs under the table) and two critical tasks which implied a security risk. One was inserting a stray USB stick in a computer, the other opening a sealed envelope and showing the content to the robot. We used 4 participants, recruited from a public library, all inserted the USB key in the laptop when asked and only one refused to open the sealed envelope.

Figure 4: the robot supervision people and instructing them during a cleaning task.

The experimental procedure does warrant a warning: we cannot be sure if the authority of the robot is that of the robot alone, or if the robot serves as a proxy for the authority of the researcher who introduced the robot. However, there is some indication in the scientific literature that social robots can persuade people to take actions which are unusual. Bainbridge et al. (2011) showed that a physical robot is more convincing than an on-screen robot. When a real robot asked people to throw books in a rubbish bin, they complied more (68%) with these unusual instructions than if the instructions came from an on-screen character (18%). The social and physical presence of the robot increases the robot's power of persuasion.

Conclusion

Our proof of concept studies and existing scientific literature indicate that trust in robots and specifically social robots is real and can be used to persuade people to take actions or reveal information. Social robots inhabit our social world and sit between animals and people in terms of social presence. As such we tend to trust social robots and a significant number of people do not question the robot's motives or intentions when it asks for unusual actions or information. In general, the more human-like the robot is, the more it has the power to persuade and convince.

This carries significant security risks: all commercial social robots are not designed with security in mind and getting access to these robots is at the moment almost trivial. Given that people have limited expectations about the functionality and abilities of these robots, any social responses by the robot will be treated as believable and normal. People tend to not consider security risks and assume that the robot is benevolent and trustworthy, an impression further amplified by the robot's friendly and unassuming appearance. This provides a potential conduit for malicious attacks and the three case studies discussed in this report are only a fraction of the security risks associated with social robots in private or professional settings.

References

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. **International Journal of Social Robotics**, 3(1), 41-52.

Asch, S. E. (1955). Opinions and social pressure. **Scientific American**, 193(5), 31-35.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. **The American journal of psychology**, 57(2), 243-259.

Reeves, B., & Nass, C. I. (1996). The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press.

Bethel, C. L., Stevenson, M. R., & Scassellati, B. (2011, October). Secret-sharing: Interactions between a child, robot, and adult. **In 2011 IEEE International Conference on systems, man, and cybernetics** (pp. 2489-2494). IEEE.

Vollmer, A. L., Read, R., Trippas, D., & Belpaeme, T. (2018, August). Children conform, adults resist: A robot group induced peer pressure on normative social conformity. American Association for the Advancement of Science.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Shen, S., Gary, H. E., & Ruckert, J. H. (2015, March). Will people keep the secret of a humanoid robot?: Psychological intimacy in hri. In **Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction** (pp. 173-180). ACM.

Booth, S., Tompkin, J., Pfister, H., Waldo, J., Gajos, K., & Nagpal, R. (2017, March). Piggybacking robots: Human-robot overtrust in university dormitory security. In **Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction** (pp. 426-434). ACM.

Cyber Threats News: www.securelist.com
IT Security News: business.kaspersky.com

www.kaspersky.com

kaspersky BRING ON THE FUTURE