Inevitable Psychological Mechanisms Triggered by Robot Appearance: Morality Included?

Bertram F. Malle
Brown University
Department of Cognitive, Linguistic,
and Psychological Sciences
Providence, RI 02906
bfmalle@brown.edu

Matthias Scheutz
Tufts University
Department of Computer Science
Medford, MA 02155
matthias.scheutz@tufts.edu

Abstract

Certain stimuli in the environment reliably, and perhaps inevitably, trigger human cognitive and behavioral responses. We suggest that the presence of such "trigger stimuli" in modern robots can have disconcerting consequences. We provide one new example of such consequences: a reversal of a pattern of moral judgments people make about robots, depending on whether they view a "mechanical" or a "humanoid" robot.

Introduction

A goose that sees a hawk-shaped silhouette in the sky will protect itself, responding inevitably to a reliable indicator of threat (Lorenz and Tinbergen 1938). Infants who observe an object that has eyes will follow its "gaze," responding inevitably to a reliable indicator of mental agency (Johnson, Slaughter, and Carey 1998). Adults who look at two lines angled toward each other will experience an object as larger when it is placed closer to the narrow end of the two lines than when it is placed closer to the wide end of the two lines (Müller-Lyer 1889). (See Figure 1.)

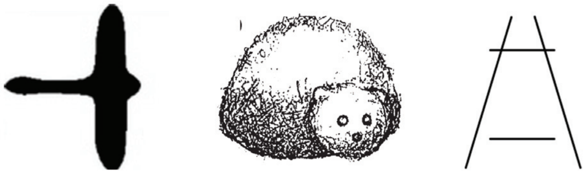

Figure 1. Examples of "Trigger Stimuli": for Geese's Protective Behavior, Infants' Agency Perception, and Adults' Line Length Estimation (From Left to Right)

These and many more examples illustrate that certain stimuli can trigger inborn or early-learned cognitive and behavioral responses that are nearly impossible to avoid. Such "trigger stimuli" often occur in nature (such as hawks or eyes), but human culture and technology have replicated many of them, thereby allowing artifacts to serve as triggers of those very same unavoidable responses—such as furry balls with glass eyes in experiments, perspectival drawings, and the entire animation industry.

Likewise for the robotics industry. Robots are artifacts with an unprecedented potential to similarly trigger such unavoidable responses: through their appearance, movements, voice quality, facial expressions, and communicative signals. Indeed, converging evidence in human-robot interaction (HRI) research points to many such triggering effects, and as robot design becomes more sophisticated, many more of these effects will be documented.

Effects of Humanlike Appearance

The most powerful stimulus in robotics is not a single feature but rather a nexus of features: human-like appearance (Złotowski et al. 2014). The co-presence of trigger stimuli such a limbs, head, eyes, facial features, etc. will lead to a wide array of inferences about a humanoid robot's "capacities"—as more intelligent, more autonomous, and as having more mind (Eyssel et al. 2012; Broadbent et al. 2013; Bartneck et al. 2009; Walters et al. 2009). Humanoid robots can also elicit ingroup bias (Eyssel and Kuchenbrandt 2012), cheater detection (Litoiu et al. 2015), and spontaneous visual perspective taking (Zhao, Cusimano, and Malle 2016). There are well-known boundaries to these kinds of inferences—the "uncanny valley" of discomfort with overly humanlike robots that may stem from contradictory evidence: some trigger stimuli in the robot promise properties that other observed behaviors clearly deny (Kätsyri et al. 2015).

Despite these boundaries, however, the array of "inevitable" reactions humans have to robots is wide and only partially understood. An unconcerned approach to robot design would simply collect these reactions as they emerge from robots being distributed in society, and add them to the list of curious human responses to technology. We believe such a wait-and-see approach would be irresponsible. Some of the inevitable responses that humanoid robots elicit in humans are costly and consequential. The emotions humans feel when seeing certain facial expressions can make them vulnerable to manipulation; the attachments humans form when being promised loyalty and affection by a robot can make the person vulnerable to loss and grieving; and the trust in a machine that appears intelligent and autonomous may be shattered in a dangerous situation outside the robot's programmed scope of action.

Moral Judgments of Robot Behavior

Morality is a consequential domain of human thought and action that is beginning to receive attention by the HRI community. It may even be the most consequential domain yet for the impact of appearance on triggered psychological responses. Human communities rely on its members to learn, apply, and enforce moral norms so that mutual trust and predictability can secure cooperative interpersonal behavior (Joyce 2006). If (or when) robots become more involved in human communities, moral norms will apply to robots as well; and so may numerous associated social practices of admonishing, blaming, apologizing, and forgiving (Malle 2015). But will the look or sound of robots influence the way humans treat robots in moral terms? Will different norms apply to human-looking robots? Will they be punished more or less than other robots? We are concerned that this may well be the case if we do not actively anticipate or perhaps intervene in the potential triggering effects of humanlike appearance.

There is growing evidence that people are quite willing to extend moral judgments to a robot (Briggs and Scheutz 2014; Kahn, Jr. et al. 2012; Monroe, Dillon, and Malle 2014). Recently we found the strongest evidence to date for humans to systematically blame robots as a function of actions and mental states in much the same way as they blame humans (Malle et al. 2015). The situation involved a moral dilemma modeled after the classic trolley dilemma (Thomson 1985; Greene et al. 2001). Such dilemmas typically involve a conflict between obeying a prosocial obligation (e.g., saving people who are in danger) and obeying a prohibition against harm (e.g., killing a person in the attempt to save those in danger). In our experiments, people held the robot responsible for its decision to resolve the dilemma, whether it decided to intervene and sacrifice one for the good of many or decided not to intervene and thereby allow four people to die. However, the data also revealed an asymmetry in people's judgments of human and robot agent. People considered a human agent's intervention (i.e., sacrificing one life while saving four lives) more blameworthy than a nonintervention; conversely, they considered a *robot* agent's nonintervention more blameworthy than an intervention.

In yet more recent experiments (Malle et al. 2016) we examined the role of robot appearance in this moral judgment context. In the initial studies we had introduced the robot only verbally, as an "advanced state-of-the art repair robot" (and compared it to a human repairman in exactly the same situation). To clarify what mental model people actually had of this "robot," our latest experiments presented people, in a between-subjects design, with either a mechanical robot or a humanoid robot (see Figure 2) and used *identical* verbal descriptions of the robot, the dilemma situation, and the robot's decision. Something rather remarkable happened: People showed the same human-robot asymmetry in their blame judgments as in the previous (text-based) experiments when they viewed the *mechanical* robot, but this asymmetry disappeared when they viewed the *humanoid* robot. The latter elicited a pattern of blame judgments parallel to that for humans: more blame for intervention than nonintervention.

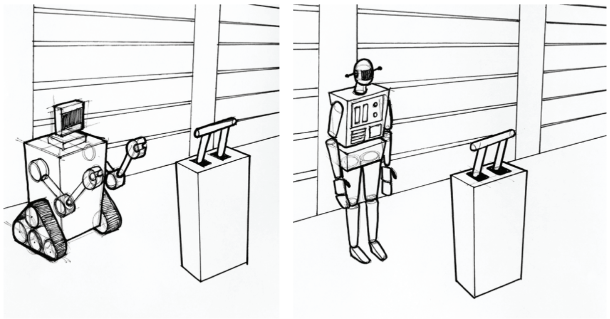

Figure 2 Illustrations of a Mechanical Robot and a Humanoid Robot as Used in a Moral Judgment Experiment

Whichever decision (intervention or nonintervention) one might favor in such a moral dilemma, it seems disconcerting that a mere drawing of a humanoid (vs. a mechanical) robot is sufficient to reverse a pattern of blame judgments about a life-and-death decision. Apparently, the mental model activated by one robot (a mechanical one) maintains a distinction in people's minds between human agent and robot agent; but the mental model activated by another (a humanoid) robot somehow eliminates this distinction. We don't know yet the exact psychological mechanisms that make people reluctant to intervene (for one hypothesis, see Greene et al. 2001) and the mechanisms that make people blame an intervening human agent more strongly; nor do we know the exact mechanisms that make people switch their moral assessments in the presence of a simple illustration. We also don't know how robust these variations are, say, in the face of life-size robots or in response to explicit instructions about the mechanical nature of the humanoid robot despite its deceptive appearance. But we better find out, through systematic empirical research, so we can provide robot designers with the necessary insights to get a handle on a potential appearance-morality confusion.

Acknowledgements

This project was supported in part by a grant from the Office of Naval Research, No. N00014-13-1-0269. The opinions expressed here are our own and do not necessarily reflect the views of ONR

References

Bartneck, Christoph, Takayuki Kanda, Omar Mubin, and Abdullah Al Mahmud. 2009. "Does the Design of a Robot Influence Its Animacy and Perceived Intelligence?" *International Journal of Social Robotics* 1 (2): 195–204. doi:10.1007/s12369-009-0013-7.

Briggs, Gordon, and Matthias Scheutz. 2014. "How Robots Can Affect Human Behavior: Investigating the Effects of Robotic Displays of Protest and Distress." *International Journal of Social Robotics* 6 (2): 1–13.

Broadbent, Elizabeth, Vinayak Kumar, Xingyan Li, John Sollers 3rd, Rebecca Q. Stafford, Bruce A. MacDonald, and Daniel M. Wegner. 2013. "Robots with Display Screens: A Robot with a More Humanlike Face Display Is Perceived to Have More Mind and a Better Personality." *PLoS ONE* 8 (8): e72589. doi:10.1371/journal.pone.0072589.

Eyssel, Friederike, D. Kuchenbrandt, S. Bobinger, L. de Ruiter, and F. Hegel. 2012. "'If You Sound like Me, You Must Be More Human': On the Interplay of Robot and User Features on Human-Robot Acceptance and Anthropomorphism." *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction, HRI'12*, March, 125–26.

Eyssel, Friederike, and Dieta Kuchenbrandt. 2012. "Social Categorization of Social Robots: Anthropomorphism as a Function of Robot Group Membership." *British Journal of Social Psychology* 51 (4): 724–31. doi:10.1111/j.2044-8309.2011.02082.x.

Greene, Joshua D., R. B. Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293 (5537): 2105–8. doi:10.1126/science.1062872.

Johnson, Susan C., Virginia Slaughter, and Susan Carey. 1998. "Whose Gaze Will Infants Follow? The Elicitation of Gaze-Following in 12-Month-Olds." *Developmental Science* 1 (October): 233–38. doi:10.1111/1467-7687.00036.

Joyce, Richard. 2006. *The Evolution of Morality*. MIT Press.

Kahn, Jr., Peter H., Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Jolina H. Ruckert, Solace Shen, Heather E. Gary, Aimee L. Reichert, Nathan G. Freier, and Rachel L. Severson. 2012. "Do People Hold a Humanoid Robot Morally Accountable for the Harm It Causes?" In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 33–40. New York, NY: ACM. doi:10.1145/2157689.2157696.

Kätsyri, Jari, Klaus Förger, Meeri Mäkäräinen, and Tapio Takala. 2015. "A Review of Empirical Evidence on Different Uncanny Valley Hypotheses: Support for Perceptual Mismatch as One Road to the Valley of Eeriness." *Frontiers in Psychology* 6: 390. doi:10.3389/fpsyg.2015.00390.

Litoiu, Alexandru, Daniel Ullman, Jason Kim, and Brian Scassellati. 2015. "Evidence That Robots Trigger a Cheating Detector in Humans." In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '15*, 165–72. http://doi.acm.org/10.1145/2696454.2696456.

Lorenz, Konrad, and Niko Tinbergen. 1938. "Taxis Und Instinkthandlung in Der Einrollbewegung Der Graugans I." *Zeitschrift Für Tierpsychologie* 2: 1–29.

Malle, Bertram F. 2015. "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology*, July. doi:10.1007/s10676-015-9367-8.

Malle, Bertram F., Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. "Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents." In *HRI '15: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction.*, 117–24. New York, NY: ACM.

Malle, Bertram F., Matthias Scheutz, Jodi Forlizzi, and John Voiklis. 2016. "Which Robot Am I Thinking about? The Impact of Action and Appearance on People's Evaluations of a Moral Robot." In *Paper Presented at HRI'16: The Eleventh Annual Meeting of the IEEE Conference on Human-Robot Interaction*. Christchruch, New Zealand.

Monroe, Andrew E., Kyle D. Dillon, and Bertram F. Malle. 2014. "Bringing Free Will down to Earth: People's Psychological Concept of Free Will and Its Role in Moral Judgment." *Consciousness and Cognition* 27 (July): 100–108. doi:10.1016/j.concog.2014.04.011.

Müller-Lyer, F. C. 1889. "Optische Urteilstäuschungen." *Archiv Für Anatomie Und Physiologie* 2: 263–70.

Thomson, Judith Jarvis. 1985. "The Trolley Problem." *The Yale Law Journal* 94 (6): 1395–1415.

Walters, M. L., K. L. Koay, D. S. Syrdal, Kerstin Dautenhahn, and R. Te Boekhorst. 2009. "Preferences and Perceptions of Robot Appearance and Embodiment in Human-Robot Interaction Trials." In *Proceedings of New Frontiers in Human-Robot Interaction*, 136–43.

Zhao, Xuan, Corey Cusimano, and Bertram F. Malle. 2016. "Do People Spontaneously Take a Robot's Visual Perspective?" In *Paper Presented at HRI '16: The Eleventh Annual ACM/IEEE International Conference on Human-Robot Interaction*. Christchurch, New Zealand.

Złotowski, Jakub, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2014. "Anthropomorphism: Opportunities and Challenges in Human–robot Interaction." *International Journal of Social Robotics*, November, 1–14. doi:10.1007/s12369-014-0267-6.