12-12-2023

Human-Robot Interactions: Insights from Experimental and Evolutionary Social Sciences

Eric Schniter

Human-Robot Interactions: Insights from Experimental and Evolutionary Social Sciences

Comments

ESI Working Paper 23-14

Human-Robot Interactions: Insights from Experimental and Evolutionary Social Sciences

by Eric Schniter

Working paper *prepared for* B. Kebede (Ed.) *Encyclopedia of Experimental Social Science*. Elgar Encyclopedias in the Social Sciences.

December 12, 2023

Experimental research in the realm of human-robot interactions has focused on the behavioral and psychological influences affecting human interaction and cooperation with robots. A robot is loosely defined as a device designed to perform agentic tasks autonomously or under remote control, often replicating or assisting human actions. Robots can vary widely in form, ranging from simple assembly line machines performing repetitive actions to advanced systems with no moving parts but with artificial intelligence (AI) capable of learning, problem-solving, communicating, and adapting to diverse environments and human interactions. Applications of experimental human-robot interaction research include the design, development, and implementation of robotic technologies that better align with human preferences, behaviors, and societal needs. As such, a central goal of experimental research on human-robot interactions is to better understand how trust is developed and maintained. A number of studies suggest that humans trust and act toward robots as they do towards humans, applying social norms and inferring agentic intent (Rai and Diermeier, 2015). While many robots are harmless and even helpful, some robots may reduce their human partner's wages, security, or welfare and should not be trusted (Taddeo, McCutcheon and Floridi, 2019; Acemoglu and Restrepo, 2020; Alekseev, 2020). For example, more than half of all internet traffic is generated by bots, the majority of which are "bad bots" (Imperva, 2016). Despite the hazards, robotic technologies are already transforming our everyday lives and finding their way into important domains such as healthcare, transportation, manufacturing, customer service, education, and disaster relief (Meyerson *et al.*, 2023).

Experiments focused on human behavior within the realm of human-robot interactions, often center on understanding how individuals make decisions and establish trust when interacting with robots (Hancock *et al.*, 2011; Hoff and Bashir, 2015). The interdisciplinary field studying human-robot interaction draws insights from evolutionary social science, behavioral economics, and robotics to illuminate the design of cognitive processes influencing human behavior toward robots. Evolutionary social science comprises an interdisciplinary field integrating evolutionary principles from psychology, anthropology, and biology to understand how the evolved design of the human mind gives rise to patterns of cognition, behavior, and culture when interacting with features of the world. Behavioral economists often rely on predictions generated by social scientists and the controlled methods of experimental economics, inducing value in the laboratory with incentive-compatible tasks to study how monetary costs and rewards, social preferences, and experiences influence decision-making in social interactions (Smith, 1976).

By applying an evolutionary perspective and conducting incentive-compatible and realistic experiments into trust-based human-robot interactions, researchers have come to a better understanding of human decision-making processes and the evolutionary design of behavioral algorithms relevant to human-robot interaction (Tooby, 2009). The study of behavioral algorithms extends to the robot algorithms –many of which are "black boxes"—that shape human behavior and are ubiquitous in our daily activities (Ishowo-Oloko *et al.*, 2019). One key take-away of this research is that people are more likely to interact with robots that they can understand and trust (Lee and See, 2004; Schaefer *et al.*, 2016). However, a psychology that evolved to navigate interactions with fellow humans regulates this trust – presenting a hazard when regulating trust in robots. This feature of our evolved psychology also fuels our fears of becoming overly-reliant on robots (Salem *et al.*, 2015; Robinette *et al.*, 2016a). But just as over-trusting harmful robots is a problem, so is under-trusting the robots that could help us (Ishowo-Oloko *et al.*, 2019). The research we review below features both undertrust and overtrust of robots.

Previous work shows that emotions regulate social behavior and may, likewise, regulate human-robot interactions. A set of basic social emotions such as guilt, gratitude, anger, and pride result from trust-based interactions and can affect how we treat others and how others treat us in trust-based interactions (Schniter and Shields, 2013; Schniter, Sheremeta and Shields, 2015). For example, a trustor's guilt—triggered when one does not trust a trustee—promotes the subsequent extension of trust, and a trustor's anger—triggered when a trustee fails to reciprocate—down-regulates trust re-extension (Schniter and Sheremeta, 2014). The degree to which social emotions may also affect trust-based interactions with robots, depends in part on whether robots are responsive to the recalibrational effects of emotions directed at them. Historically, humans have recognized robots as lacking social presence: unable to understand, respond to, or experience social emotions (Gray, Gray and Wegner, 2007; Spence *et al.*, 2014). More recently, communicative companionship robots and robo-pets have been developed that not only produce emotional expressions, but respond to human emotional needs, providing a therapeutic relief from loneliness (Austermann *et al.*, 2010; Broadbent, 2017).

Trust in robots may also depend on how our interactions with robots affect other people, how robots affect us in survival situations, and how intuitively we understand the goals or intentions behind robot behavior. Below, we review three domains of research that have been productive in unraveling these concerns affecting human-robot interactions: experiments with economic game interactions, emergency and warfare situations, and autonomous vehicle interactions.

Human-robot interactions have been studied in the context of human–robot Prisoner Dilemma (PD) game experiments (Hsieh, Chaudhury and Cross, 2023; Maggioni and Rossignoli, 2023). In the repeated PD with robots, participants show an initially high rates of cooperation -similar to that shown when interacting in the PD with fellow humans. The initially high level of cooperation with robots quickly decays after the game is repeated, giving way to a tit-for-tat response strategy.

Despite the robot having played a purely random strategy in the Hsieh et al. (2023) study, many participants incorrectly interpreted that the robot had played reciprocally. The Coin Exchange game, a hybrid of the PD and the Trust game, has also been studied in the context of experimental human–robot game interactions (Wu *et al.*, 2016). Wu and colleagues (2016) showed that participants initially trusted robots more than humans and cooperated with a human opponent just as readily as a robot opponent. Behavioral economists have also used trust game experiments to study trust and emotional reactions to trust-based interactions with fellow humans and with robots programmed to mimic human reciprocity (Schniter, Shields and Sznycer, 2020). In these interactions, people extend trust similarly to fellow humans and to robots programmed to act like humans, but experienced emotions differently with respect to robot versus human partners. For example, a trustee's failure to reciprocate the trustor's trust triggered more trustor anger when the trustee was a human and less trustor anger when the trustee was a robot. Similarly, reciprocation by trustees triggered more trustor gratitude when the trustee was a human and less trustor gratitude when the trustee was a robot. Further, human trustors' emotions finely discriminated among robot types. Human trustors experienced more pride and less guilt upon placing trust in a trustee. But pride and guilt were more intense when the trustee was a robot whose payoffs went to a fellow human than when the trustee was a robot acting alone. Interestingly, the differences in emotional experience across partner types appear to be restricted to the domain of social emotions and do not generalize to the domain of non-social emotions.

Researchers have leveraged virtual reality technology in the laboratory to simulate real-world crises, for example emergency fire and active shooter situations, and to study human trust in robots offering evacuation assistance and giving advice about whom to kill or not during warfare (Holbrook, et al., 2023a, 2023b; Robinette et al., 2016). Across these scenarios, humans overly trust robots. Survival and warfare scenarios feature the potential for extreme loss of life and have three features that make them ideal for an error-management heuristic that will err to self-preservation and cautious trust in order to make the least costly error (Haselton and Nettle, 2005): (a) decisions in these scenarios are based on uncertain information, and over the course of human evolutionary history, (b) decisions in these scenarios had recurrent impacts on survival and genetic fitness, and (c) the costs of false-positive and false-negative errors associated with these decisions have recurrently been asymmetrical.

Another area of human-robot interaction where life is often on the line is humans' use of autonomously driven vehicles. When people can make choices about how to program autonomous vehicles, their interactions with others in traffic are more cautious and cooperative than if they were driving themselves (de Melo, Marsella and Gratch, 2019). This can be explained by humans showing preference for automated driving programs that err to caution, contrasted with the human driver who occasionally throws caution to the wind when running late, enraged by other drivers, or otherwise distracted. Travelling in motorized vehicles is an evolutionarily novel activity that modern humans show remarkable ease engaging in on a daily basis, despite fatal motor vehicle

collisions (2.5% of all deaths) ranking first place among non-disease related reasons for death (Dattani *et al.*, 2023) and leading as the primary cause of injury and death among children worldwide (World Health Organization, 2008). Unfortunately, people are hesitant to delegate driving responsibility to autonomous vehicles and observers critically judge such delegations as shirking moral responsibility (Gogoll and Uhl, 2018).

In summary, the experimental social science research on human-robot interactions focuses on unraveling decision-making processes underlying human relationships and the relationships humans are constantly developing with robots. Evolutionary theory serves as a foundational framework for understanding the adaptive functions and evolutionary origins of human behaviors relevant to human-robot interaction and experimentally studied by behavioral economists. According to this perspective, the architecture of the human mind evolved to have enough structure and content to promote our ancestors' survival and reproduction while also having the flexibility to navigate novel challenges and opportunities (Barkow, Cosmides and Tooby, 1992). These features enable humans to design and rationally interact with AI and robots—agents whom our forager ancestors may never have imagined could exist. Still, interactions with robots, and science's ability to explain these interactions, are imperfect, because robots (i) lack the psychophysical cues that we evolved to expect in an interaction partners and (ii) often are guided by unexplainable or unintuitive decision logics. Future relationships with robots will depend on the creation of and trust in robots that effectively engage with humans across diverse contexts, fostering trust, cooperation, and successful interactions that align with human cognitive mechanisms, social tendencies, and cultural diversity. Our abilities to design, build, and interact with robots are testament to the power of human cognition. A behavioral science of human social interactions with AI and robots that harnesses this power has great promise.

References

Acemoglu, D. and Restrepo, P. (2020) 'Robots and Jobs: Evidence from US Labor Markets', *Journal of Political Economy*, 128(6), pp. 2188–2244. Available at: https://doi.org/10.1086/705716.

Alekseev, A. (2020) 'The Economics of Babysitting a Robot', *ESI Working Papers* [Preprint]. Available at: https://digitalcommons.chapman.edu/esi_working_papers/324.

Austermann, A. *et al.* (2010) 'How do users interact with a pet-robot and a humanoid', in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery (CHI EA '10), pp. 3727–3732. Available at: https://doi.org/10.1145/1753846.1754046.

Barkow, J.H., Cosmides, L. and Tooby, J. (1992) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press.

Broadbent, E. (2017) 'Interactions with robots: The truths we reveal about ourselves', *Annual review of psychology*, 68, pp. 627–652.

Dattani, S. *et al.* (2023) ‘Causes of Death’, *Our World in Data* [Preprint]. Available at: https://ourworldindata.org/causes-of-death (Accessed: 11 December 2023).

Gogoll, J. and Uhl, M. (2018) ‘Rage against the machine: Automation in the moral domain’, *Journal of Behavioral and Experimental Economics*, 74, pp. 97–103. Available at: https://doi.org/10.1016/j.socec.2018.04.003.

Gray, H.M., Gray, K. and Wegner, D.M. (2007) ‘Dimensions of Mind Perception’, *Science*, 315(5812), pp. 619–619. Available at: https://doi.org/10.1126/science.1134475.

Hancock, P.A. *et al.* (2011) ‘A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction’, *Human Factors*, 53(5), pp. 517–527. Available at: https://doi.org/10.1177/0018720811417254.

Haselton, M.G. and Nettle, D. (2005) ‘The paranoid optimist: an integrative evolutionary model of cognitive biases.’

Hoff, K.A. and Bashir, M. (2015) ‘Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust’, *Human Factors*, 57(3), pp. 407–434. Available at: https://doi.org/10.1177/0018720814547570.

Holbrook, C., Holman, D., Wagner, A.R., *et al.* (2023) ‘Investigating Human-Robot Overtrust During Crises’. PsyArXiv. Available at: https://doi.org/10.31234/osf.io/3p5qe.

Holbrook, C., Holman, D., Clingo, J., *et al.* (2023) ‘Overtrust in AI Recommendations to Kill’. Available at: https://doi.org/10.31234/osf.io/2umct.

Hsieh, T.-Y., Chaudhury, B. and Cross, E.S. (2023) ‘Human–Robot Cooperation in Economic Games: People Show Strong Reciprocity but Conditional Prosociality Toward Robots’, *International Journal of Social Robotics*, 15(5), pp. 791–805. Available at: https://doi.org/10.1007/s12369-023-00981-7.

Imperva (2016) *Bot Traffic Report 2016 | Imperva*. Available at: https://www.imperva.com/blog/bot-traffic-report-2016/ (Accessed: 11 December 2019).

Ishowo-Oloko, F. *et al.* (2019) ‘Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation’, *Nature Machine Intelligence*, 1(11), pp. 517–521. Available at: https://doi.org/10.1038/s42256-019-0113-5.

Lee, J.D. and See, K.A. (2004) ‘Trust in Automation: Designing for Appropriate Reliance’, *Human Factors*, 46(1), pp. 50–80. Available at: https://doi.org/10.1518/hfes.46.1.50_30392.

Maggioni, M.A. and Rossignoli, D. (2023) ‘If it looks like a human and speaks like a human ... Communication and cooperation in strategic Human–Robot interactions’, *Journal of Behavioral and Experimental Economics*, 104, p. 102011. Available at: https://doi.org/10.1016/j.socec.2023.102011.

de Melo, C.M., Marsella, S. and Gratch, J. (2019) ‘Human Cooperation When Acting Through Autonomous Machines’, *Proceedings of the National Academy of Sciences*, 116(9), pp. 3482–3487. Available at: https://doi.org/10.1073/pnas.1817656116.

Meyerson, H. *et al.* (2023) ‘Introductory Chapter: Human-Robot Interaction–Advances and Applications’, in *Human-Robot Interaction. Edited by Ramana Vinjamuri*. IntechOpen. Available at: https://mdsoar.org/bitstream/handle/11603/28105/85438.pdf?sequence=3 (Accessed: 22 October 2023).

Rai, T.S. and Diermeier, D. (2015) ‘Corporations are Cyborgs: Organizations elicit anger but not sympathy when they can think but cannot feel’, *Organizational Behavior and Human Decision Processes*, 126, pp. 18–26. Available at: https://doi.org/10.1016/j.obhdp.2014.10.001.

Robinette, P. *et al.* (2016a) ‘Overtrust of Robots in Emergency Evacuation Scenarios’, in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. Piscataway, NJ, USA: IEEE Press (HRI ’16), pp. 101–108. Available at: http://dl.acm.org/citation.cfm?id=2906831.2906851 (Accessed: 5 April 2019).

Robinette, P. *et al.* (2016b) ‘Overtrust of robots in emergency evacuation scenarios’, in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 101–108. Available at: https://doi.org/10.1109/HRI.2016.7451740.

Salem, M. *et al.* (2015) ‘Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust’, in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: ACM (HRI ’15), pp. 141–148. Available at: https://doi.org/10.1145/2696454.2696497.

Schaefer, K.E. *et al.* (2016) ‘A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems’, *Human Factors*, 58(3), pp. 377–400. Available at: https://doi.org/10.1177/0018720816634228.

Schniter, E. and Sheremeta, R.M. (2014) ‘Predictable and predictive emotions: explaining cheap signals and trust re-extension’, *Frontiers in Behavioral Neuroscience*, 8. Available at: https://doi.org/10.3389/fnbeh.2014.00401.

Schniter, E., Sheremeta, R.M. and Shields, T.W. (2015) ‘Conflicted emotions following trust-based interaction’, *Journal of Economic Psychology*, 51, pp. 48–65. Available at: https://doi.org/10.1016/j.joep.2015.08.006.

Schniter, E. and Shields, T. (2013) ‘Recalibrational emotions and the regulation of trust-based behaviors’, in *Psychology of trust: New research*. Hauppauge, NY, US: Nova Science Publishers (Psychology of emotions, motivations and actions), pp. 1–58.

Schniter, E., Shields, T.W. and Sznycer, D. (2020) ‘Trust in humans and robots: Economically similar but emotionally different’, *Journal of Economic Psychology*, 78, p. 102253. Available at: https://doi.org/10.1016/j.joep.2020.102253.

Smith, V.L. (1976) ‘Experimental Economics: Induced Value Theory’, *The American Economic Review*, 66(2), pp. 274–279.

Spence, P.R. *et al.* (2014) ‘Welcoming Our Robot Overlords: Initial Expectations About Interaction With a Robot’, *Communication Research Reports*, 31(3), pp. 272–280. Available at: https://doi.org/10.1080/08824096.2014.924337.

Taddeo, M., McCutcheon, T. and Floridi, L. (2019) ‘Trusting artificial intelligence in cybersecurity is a double-edged sword’, *Nature Machine Intelligence*, 1(12), pp. 557–560. Available at: https://doi.org/10.1038/s42256-019-0109-1.

Tooby, J. (2009) *The Great Pivot: Artificial Intelligences, Native Intelligences, and the Bridge Between.*, *Edge.org John Brockman, Editor and Publisher*. Available at: https://www.edge.org/response-detail/11265 (Accessed: 11 December 2023).

World Health Organization, (WHO) (2008) 'World report on child injury prevention. 2008'. Available at: http://www.who.int/violence_injury_prevention/child/injury/world_report/report/en/index.html (Accessed: 11 December 2023).

Wu, J. *et al.* (2016) 'Trust and Cooperation in Human-Robot Decision Making', in *2016 AAAI Fall Symposium Series*. *2016 AAAI Fall Symposium Series*. Available at: https://www.aaai.org/ocs/index.php/FSS/FSS16/paper/view/14118 (Accessed: 5 April 2019).