Reconsidering Deception in Social Robotics: The Role of Human Vulnerability (Student Abstract)

Rachele Carli[1, 2], **Amro Najjar**[3]
[1]Alma AI, University of Bologna
[2] ICR, University of Luxembourg
[3] LIST Institute, University of Luxembourg
rachele.carli2@unibo.it, amro.najjar@list.lu

Abstract

The literature on deception in human-robot interaction (henceforth HRI) could be divided between: (i) those who consider it essential to maximise users' end utility and robotic performance; (ii) those who consider it unethical, because it is potentially dangerous for individuals' psychological integrity. However, it has now been proven that humans are naturally prone to anthropomorphism and emotional attachment to inanimate objects. Consequently, despite ethical concerns, the argument for the total elimination of deception could reveal to be a pointless exercise. Rather, it is suggested here to conceive deception in HRI as a dynamic to be modulated and graded, in order to both promote innovation and protect fundamental human rights. To this end, the concept of vulnerability could serve as an objective balancing criterion.

Introduction

The historical diatribe around deception in HRI can be summarised in two main approaches: (i) deception as an essential element in the design and operation of some devices – to enhance acceptability, collaboration, efficiency in performances (Isaac and Bridewell 2017); (ii) deception as substantially unethical, because it is potentially dangerous for individuals' integrity (Sharkey and Sharkey 2021). Added to this, numerous studies in cognitivism and neuroscience have underlined that humans are naturally inclined to create affective bonds with new technologies (Chaminade, Hodgins, and Kawato 2007). Therefore, an approach that only aims to enhance or banish such dynamics in HRI proves ineffective. Instead, it is here suggested that deception could be conceived as a paradigm to be addressed and modulated with due regard for both technical requirements and respect for fundamental rights. The main criterion around which to build this balance could be that of human vulnerability.

The Centrality of Deception in HRI

Deception is a central element in AI systems, as amply demonstrated since the Turing Test. Even not aiming to make the user believe that the robotic counterpart is – in fact – animated, deception is considered functional to convey an experience of sociability, that enhances engagement with the device. For this specific purpose, robots are programmed so as to: appear in need of support – encouraging teamwork (Budde et al. 2018); seem fallible and clumsy – eliciting empathy and protective instincts (Lammer et al. 2011); utter reassuring phrases that reveal emotional participation – conveying trust and familiarity (Natale 2020); emulate the human need to reflect before acting – giving the idea of having a personality and an intentionality, thus covering technical inefficiencies. Moreover, in rescue or care contexts, deception is considered crucial for the robot to be able to deal with anxious or not fully conscious individuals, and to appear better able to cope with contingencies (Shim and Arkin 2015). Physical design is also essential to make the user feel comfortable and ready to increase interaction.

This quick review of how deception is perpetrated in HRI demonstrates that it could be useful to: (i) convey positive feelings towards the machine, which can thus work more efficiently, (ii) minimise malfunctions, and (iii) make the device capable to act in a human-centred context. This, however, leaves many ethical and legal questions open.

The Risks of Deception for the Final User

What is emerging is that deceptive dynamics create a mere appearance – of sociability, of caring, of empathic response – that does not reflect what machines are actually capable of doing, yet. A branch of research severely criticises this, emphasising the potential harmfulness and the concerning ethical-legal implications of this theme.

The proponents of this theory claim for the right of individuals ”to see reality for what it really is” (Sparrow 2002). Conversely, the deceptive phenomenon is perceived as manipulative, capable of distorting people's perception of themselves and the world. One of the main fears is the emergence of ”mechanomorphism” (Caporael 1986), thus inducing humans to adapt their expectations, relational dynamics, and coping strategies to the capabilities of the machine, rather than the opposite. Some of the extreme consequences could be dehumanisation of care, loss of meaningful human contacts, and threat to personal data.

Anthropomorphism and Humanity

From the Media Equation Theory onward, it has been proven that individuals are characterised by the so called ”symbolic interactionism”, namely the faculty of constructing meanings through interaction (von Scheve 2014). In an HRI scenario, anthropomorphism – the tendency to ascribe human characteristics to robots – responds to this same phenomenon. Hence, it should be considered a ”default schema” (Caporael 1986), inherent to functionality of human psyche.

If it is true that anthropomorphism in HRI could be elicited and brought to its extreme consequences by specific robotic figures, it is also true that it cannot be eliminated. Therefore, the sole attempt to remove any element of anthropomorphisation, or to prohibit some of the tactics which coveys, could turn out to be not only short-sighted, but also ineffective. Short-sighted, because it would demonstrate a wrong interpretation, even a devaluation, of human complexity. Ineffective, because it would hinder the market and the very development of technologies that could be not necessarily harmful regardless, without being able to completely eliminate our tendency to over-trust and empathise.

Vulnerability: a Balancing and Guiding Tool

It is emerging more and more clearly in the literature that the figure of the perfect rational individual”, capable of performing truly free and informed actions, is only a juridical fiction. In fact, human beings are vulnerable by nature – as the Theory of Vulnerability clearly demonstrates (Gordon-Bouvier 2021) – and cannot get rid of this universal and enduring condition. The use of care robots, for example, may address some vulnerabilities – such as the paucity of assistive resources in the face of an ageing population – but exposes to others – such as the possibility of deception. The same happened in the past with other technologies. The invention and widespread use of mobile phones has made communication easier, but it has also exposed us to a reduction and modification of the mental processes involved in storing and searching for information, on an evolutionary level (Ienca 2021). Thus, we should conclude that none can be immune from vulnerability, or made vulnerable by external elements (Coeckelbergh 2013), but there are dynamics that highlight our inherent vulnerability in a more manifest and critical way.

In this theoretical context, the State can play a central role in reducing, mitigating and counterbalancing vulnerability (Fineman 2010), through the law – the only governance instrument that can create enforceable and binding obligations. Therefore, it is here suggested to use vulnerability as an objective criterion to assess the appropriateness and impact of deception in HRI, depending on the category of users and technology involved. Hence, it will be possible to achieve a twofold result: (i) to establish *ex ante* when deception should be prohibited, because it is excessively risky for the psycho-physical integrity of the users; (ii) the degree of deception that is gradually considered tolerable, while fully respecting fundamental human rights.

Conclusions and Future Works

The paper briefly reviewed the polarised diatribe between the proponents of the benefits of deception in HRI and its detractors. Then, it has been highlighted that anthropomorphism and empathy towards inanimate agents are irrevocable human phenomena, on which a real control cannot be exercised. Thus, it has been suggested a new reading of vulnerability, to be conceived as a resource and as a useful tool to evaluate and modulate deception levels, in order to both foster innovation and protect individuals’ fundamental rights.

Future works will focus on formalising the concrete way through which deception can be graded – both at a technical and regulatory level.

References

Budde, V.; Backhaus, N.; Rosen, P. H.; and Wischniewski, S. 2018. Needy robots-designing requests for help using insights from social psychology. In *2018 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, 48–53. IEEE.

Caporael, L. R. 1986. Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in human behavior*, 2(3): 215–234.

Chaminade, T.; Hodgins, J.; and Kawato, M. 2007. Anthropomorphism influences perception of computer-animated characters’ actions. *Social cognitive and affective neuroscience*, 2(3): 206–216.

Coeckelbergh, M. 2013. *Human being@ risk: Enhancement, technology, and the evaluation of vulnerability transformations*. Springer.

Fineman, M. A. 2010. The vulnerable subject: Anchoring equality in the human condition. In *Transcending the boundaries of law*, 177–191. Routledge-Cavendish.

Gordon-Bouvier, E. 2021. The vulnerable subject: Anchoring equality in the human condition (Martha Fineman). In *Leading Works in Law and Social Justice*, 226–239. Routledge.

Ienca, M. 2021. Brain and Mental Health in the Era of Artificial Intelligence. In *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues*, 261–263. Springer.

Isaac, A.; and Bridewell, W. 2017. Why robots need to deceive (and how). *Robot ethics*, 2: 157–172.

Lammer, L.; Huber, A.; Zagler, W.; and Vincze, M. 2011. Mutual-Care: Users will love their imperfect social assistive robots. In *Work-in-progress Proceedings of the international conference on social robotics*, 24–25.

Natale, S. 2020. To believe in Siri: A critical analysis of AI voice assistants. *Communicative Figurations Working Paper*, 32.

Sharkey, A.; and Sharkey, N. 2021. We need to talk about deception in social robotics! *Ethics and Information Technology*, 23(3): 309–316.

Shim, J.; and Arkin, R. C. 2015. The benefits of robot deception in search and rescue: Computational approach for deceptive action selection via case-based reasoning. In *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, 1–8. IEEE.

Sparrow, R. 2002. The march of the robot dogs. *Ethics and information Technology*, 4(4): 305–318.

von Scheve, C. 2014. Interaction rituals with artificial companions: From media equation to emotional relationships. *Science, Technology & Innovation Studies*, 10(1): 65–83.