Do Emotions Matter in the Ethics of Human-Robot Interaction? - Artificial Empathy and Companion Robots

Bert Baumgaertner[1] and **Astrid Weiss**[2]

Abstract. In this position statement we shall argue that emotions are not directly relevant in the ethics of human-robot interaction, particularly in the context of robot care-givers and human care-receivers. Our argument is based on (1) current theories of emotion and (2) empirical findings on organizational emotion research in health care. We use a thought experiment to guide the reader through aspects of emotional empathy that support our conclusion. Our general argument is that what matters to care behavior is just the relevant behavior, not the source that drives the behavior. Our reflection will show that emotional deception may not directly impact the care-receiver (as often assumed in HRI) but more relevantly other care personnel.

1 Introduction

If we knew concretely what emotions were, we would have a better idea of how to design artificial systems that have them (or, we would be able to develop less prejudiced arguments to explain why artificial systems cannot or should not have them). Our best theories of emotions, however, are not concrete enough for this end and remain at a high level of abstraction.

Consider a key theory among the dimensional theories of emotion as an example: the two-factor theory of Schachter [12]. According to Schachter [12], feelings are caused by two independent components: physical activation and cognitive evaluation of this activation. These components then converge to form the perception of a "feeling". According to this theory, the physiological activation of a person could be interpreted by this person as a result of the cognitive evaluation process, either as positive or as negative. Whether a particular physical sensation is considered to be pleasurable or painful depends entirely on the cognitive assessment.

This view is criticized by Zajonc [21]. He claims that humans can experience emotions without cognitive interpretation. Although this theory seems intuitive at first, it cannot be the full truth in its strict form, as it is hard to imagine any mental state, except for death, where no cognition whatsoever is present. Lazarus [7] for instance stresses that even strong emotions require a small amount of cognition because without it, there is no way that we can know why we are reacting. He advanced the view that cognitive ratings of a situation could be unconscious, but that these ratings are necessary for the formation of emotions.

We claim that any theory of (human) emotions will involve a physical and a behavioral component. Human emotions have a physical component in that they have biological correlates involving the limbic system and neurochemicals such as dopamine, noradrenaline, and serotonin. Behavioral components of emotions include their assertion through facial expressions, bodily reactions, and vocalization.

In HRI research, it has become more and more evident that users should establish some kind of long-term emotional/social bonding towards a companion robot [18]. As defined by Dautenhahn and colleagues "a robot companion is a robot that (i) makes itself 'useful', i.e. is able to carry out a variety of tasks in order to assist humans, e.g. in a domestic home environment, and (ii) behaves socially, i.e. possesses social skills in order to be able to interact with people in a socially acceptable manner" [3]. However, it seems to us that, especially with respect to companion robots for older adults, the topics of emotion, artificial empathy, and questions of deception and ethics have become more prominent than the topic of usefulness [20]. For example, should we be ethically concerned about handing over elderly care to robots that "fake emotion"? Does this count as a form of deception? Could this harm older adults?

Such questions, however, can be misleading. What matters to the success of building robot companions is the relevant behaviour, not the source of that behaviour, which goes in line with the argumentation of behaviour based robotics from Arkin [1]. We suggest that this also includes the ethical dimension. Hence, we argue that unless a theory of emotions is put forward on purely behavioral grounds, a theory of emotions is unnecessary for an ethics of human-robot interaction for companion robots for older adults. Companion robots are considered to be beneficial to their users to the extent that the robots (seem to) express and recognize emotions, behave according to social norms, and establish something like a care-taker/care-receiver interaction pattern. That, we argue, is the extent to which we need an ethic for human-robot interaction – with or without the emotions.

2 Related Work

Ethical concerns related to companion robots for elderly care are gaining more and more attention (see e.g. [15] & [13]). Besides topics such as loss of privacy and reduction of care duties of humans, topics such as artificial emotions and deception are becoming more prominent. We agree with others that there is a pressing need to identify the likely effects of companion robots for the aging population before they become common place and that deception and artificial empathy are key topics to be addressed. Older adults that can no longer independently live at home without assistance need company, love, and attention. At the current stage of technological development companion robots are far from offering that in a human-like manner. However, a lot of research efforts are put in emotion recognition, interpretation, and expression. There are already care robots that express emotions via facial expressions following a care-giver and care-receiver interaction paradigm, but their actual interaction

[1] University of Idaho, Idaho USA, email: bbaum@uidaho.edu
[2] Vienna University of Technology, Austria, email: astrid.weiss@tuwien.ac.at

and communication abilities are still very limited [4].

It has been argued that any benefits of companion robots for elderly care are consequences of deceiving older adults into thinking that they could establish a relationship towards the machine over time [15]. This concern is also mentioned by Turkle [17], who claims that it leaves an uncomfortable feeling to assume that in the future grandma and grandpa will say "I love you" to a robot which returns the same phrase. She also states that for the future development of such technology we need to think about the degree of "authenticity we require of our technology". Wallach and Allen [19] also argue that the detection of human social cues and robot feedback with human-like cues are general forms of deception.

These lines of argument are red herrings. Consider the claim from Sharkey and Sharkey [13] that humans can be "all too ready to anthropomorphize machines and other objects and to imagine that they are capable of more than it is actually the case". We agree that there are such circumstances, but the extent to which they should be of concern is not with respect to deception, but with respect to limitations of the capacities of the relevant robots. Humans can choose to act as though something was real even when they know it is not, as Zizek [22] notes: "I know very well that this is just an inanimate object, but none the less I act as if I believe that this is a living being". If there are circumstances where it is rational to act in this manner (which we believe there are), then surely it can be rational to act as if something has emotions even when it does not. After all, we find it acceptable to mask our own emotions to achieve ulterior goals, such as avoiding hurting someone else's feelings or keeping an emotional distance to them. In the next section we extend our line of argument through a thought experiment.

3 Thought Experiment

Our thought experiment proceeds by considering a generic caregiving scenario with humans. Suppose Eleanor is an elderly woman in a nursing home. A younger woman, Janice, works at the nursing home and is assigned to take care of Eleanor. We now ask, are Janice's emotions relevant to an ethics of care with respect to Eleanor? Surely the answer to this question depends on Eleanor and her perceptions of Janice's care (we consider the perspective from caregivers later). So let us suppose that in one situation Eleanor is happy with her care, and in another she is not.

Consider the situation where Eleanor is unhappy about how she is being cared for. There are two plausible reasons for this. Either Janice's care behavior towards Eleanor is unsatisfactory (e.g., Janice is too rough with Eleanor from Eleanor's perspective), or it is satisfactory but there is something else about Janice that makes Eleanor unhappy. The latter case might come about because Janice, despite doing a satisfactory job, does not "actually care" about Eleanor. Should Eleanor find out about this, Janice's negative or unsympathetic emotions can work to defeat Eleanor's happiness with respect to her care.

Now consider the situation in which Eleanor is happy about how she is being taken care of by Janice. For this to happen it must be the case that Janice's care behavior towards Eleanor is more or less satisfactory. Where that behavior is less satisfactory, Eleanor may be more forgiving by knowing (for argument's sake) that Janice "actually cares" for her. It seems to us implausible, however, that good intentions and the "right" emotions can make up for any level of unsatisfactory care behavior.

In each of the four cases we outlined it was the behavior that mattered, not the drivers of that behavior (such as intentions or emotions). What this is supposed to show is that considerations about emotional behavior, at least with respect to care, *dominate* considerations about the sources of that behavior. We use this to develop the rest of our argument.

4 Reflection on Emotions and Care

If considerations about behavior dominate considerations about the sources of that behavior, then the role of emotion in developing companion robots is secondary to the role of the behavior. Flipped around, the extent to which emotions matter is given by the extent to which a person's positive perception of someone else's care behavior can be defeated by knowledge of the emotional source of that behavior. In short, there is an asymmetric relationship between behavior and emotion in care-giver ethics: behavior could make up for a lack of emotions, but a lack of the appropriate behavior cannot be exonerated by emotions (at least not, we are suggesting, in this caregiver context). If we are right about this, we think several conclusions follow.

First, emotions are unnecessary with respect to an ethics of human-robot interaction. As humans we have learned that emotional states tend to be correlated with behaviors. Of course, these correlations can be broken, as in cases where persons suppress their anger, or where persons are acting. And it is precisely because these correlations can be broken that we get a separation between the behavioral and emotional components we are considering here. Moreover, what our thought experiment is supposed to show is that we (from patient perspective) favor the behavioral part over the emotional part in a care-giver context. So at least in principle, we could get away with just the behavioral component to successfully develop an ethical companion-robot.

Our second conclusion is stronger: emotions *should* not be considered in an ethics of care with respect to companion robots. Emotions can get in the way of effective care behavior, more so than they can be of help. For example, if a health professional has been emotionally compromised then they may be deemed unprofessional to engage in care behavior [6]. Emotions such as anger, fear, rage, irritation, etc., can be dangerous in the context of taking care of others. They are not only distracting, but can also lead to malicious behavior, particularly if those emotions are targeted towards the relevant subjects. In contrast, a robot care-taker will have few (if any) biases that could get in the way of providing the necessary case. Again, it is the behaviour that matters, not the source of the behaviour.

One might object that without an "emotional drive" a robot caretaker would lack the appropriate wherewithal required for care behavior. After all, the "delicate" touch of a human tends to be so precisely because of the emotional state of the caregiver. This objection, however, concedes our point. What really matters is the "delicacy" of the care, not the emotional source of such care. We argue for this point further by taking into consideration the relegated role of emotions in health care practices.

In traditional Western medical settings care givers must align their personal experiences and emotional expressions to the organizational and occupational norms of appearing unemotional [5]. The degree of this emotional distance may vary across practitioners and organizations, but it is still a dominant strategy to keep emotional neutrality. And while there is a recent tendency to practice "feeling and showing empathy" in these professions, the adoption of this practice is not to make the care receiver feel better, but is adopted in the interest of the caregiver because emotional neutrality is very emotionally demanding [8].

More specifically, there are three main themes in organizational

emotion literature: (1) regulated emotions, (2) the influence of detached emotions on the patient, and (3) cultural forms of negotiating feelings [10]. With respect to (1), care-givers experience an emotional dissonance between their individual emotional ownership and the organizational emotional ownership. For example, doctors after death-telling (and intense emotional labor of showing empathy but not feeling personal guilt) tend to leave relatives with nurses to offer additional emotional support [2]. In regards to (2), the expression of emotional neutrality of the care personnel (which is again hard emotional labor as mentioned above) can have a social influence on the care-receivers in a way that they also feel similarly "detached" to their conditions [16]. And with respect to (3), organizational culture research demonstrates that emotional performances in care settings are often implicit informal processes which are taught by observation in symbolic activities. In other words, there is an emotional socialization of health care providers [14].

If we take into account emotion management strategies and performances in the health care sector, it is easy to see that these professions are involved in emotional role-playing. From this care-giver perspective companion robots could have a significant advantage over humans in all three aspects above. Individuals who perform emotional labor may suffer from stress or burnout and may develop a concern of being "desensitized" to their patients [9]. We concede that, if robots do not take over any of the emotional labor, this aspect could become an even bigger burden for care personnel. However, it seems like the aspect of emotional distance between the robot and the care-receiver might even have a positive aspect in terms of not perceiving the conditions as very concerning (if it is just a robot taking care of me it cannot be that critical). An open issue is how companion robots will affect the emotional socialization of care-givers. This aspect is hard to predict: Will robots serve as role model to be emotionally detach or will new strategies evolve how to regulate emotions?

A related issue we think needs to be addressed is the trust that patients may place in robots. Good care behavior and the appearance of emotions by a robot may lead patients to trust the robot beyond what it is capable of doing. It is important to recognize that this question about misplaced trust is different than the question about deception we started with at the beginning. The issue we are pointing to is not about whether the behavior and emotional drive are in sync. It is strictly an issue about behavior: How can robots be made so that their behavior reflects what they are capable of doing without inviting patients to overgeneralize? We think that, given our lack of a concrete understanding of emotions in the first place, research focused on these sorts of questions are more fruitful for an ethics of human-robot interaction than concerns about authenticity or deception.

5 Conclusion

To summarize, our suggestion is that the deceptive aspect of emotions is not crucial for an ethics of robot companions and care-takers in case of robot care-givers and human care-receivers. This is because emotions are either unnecessary entirely, or the extent to which they do play is sufficiently encompassed by the relevant behavior.

We thereby follow the conclusions of Sharkey and Sharkey [13], that considering robot companions as unethical because their effectiveness depends on deception oversimplifies the issue. If we can develop a system that effectively delivers what we deem to be appropriate care behavior, then the only source of objection - though one could hardly call it even that - would be our prejudices.

With respect to organizational emotional research in the care sector, artificial emotion recognition and expression will likely affect the other care personnel. The emotional burden may increase and emotion management and regulation may have to change. Thus, we want to encourage the HRI community to consider the impact of artificial empathy from a broader sociological perspective than just with a focus on deception of care-receivers. First empirical findings from fellow researchers also support our claim: Even a "simple delivery robot" impacts and changes the emotional and social labor in hospitals [11]. To our conviction the impact of robot care-givers on emotional labor of health practitioners is of bigger societal impact, as is the issue of how to design robot behavior that accurately reflects their capacities.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 288146 ("HOBBIT") and from the Austrian Science Foundation (FWF) under grant agreement T623-N23 ("V4HRC"). Moreover, we would like to thank Donald McKendrick for the academic match making that lead to the discussions published in this position statement.

REFERENCES

[1] Ronald C Arkin, 'Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part i: Motivation and philosophy', in *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, pp. 121–128. IEEE, (2008).

[2] Robert E Clark and Emily E LaBeff, 'Death telling: Managing the delivery of bad news', *Journal of Health and Social Behavior*, 366–380, (1982).

[3] Kerstin Dautenhahn, Sarah Woods, Christina Kaouri, Michael L Walters, Kheng Lee Koay, and Iain Werry, 'What is a robot companion-friend, assistant or butler?', in *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pp. 1192–1197. IEEE, (2005).

[4] D. Fischinger, P. Einramhof, W. Wohlkinger, K. Papoutsakis, P. Mayer, P. Panek, T. Koertner, S. Hofmann, A. Argyros, M. Vincze, A. Weiss, and C. Gisinger, 'Hobbit - the mutual care robot', in *Assistance and Service Robotics in a Human Environment Workshop in conjunction with IEEE/RSJ International Conference on Intelligent Robots and Systems*, (2013).

[5] M.S. Goldstein, 'The origins of the health movement', in *Helath, illness, and healing: Society, social context, and self*, ed., D. A. Charmaz, K. Paterniti, Roxbury, Los Angeles, (1999).

[6] Nicky James, 'Divisions of emotional labour: Disclosure and cancer', *Communication, relationships and care: A reader*, 259–269, (1993).

[7] Richard S Lazarus, 'Cognition and motivation in emotion.', *American psychologist*, **46**(4), 352, (1991).

[8] Stewart W Mercer and William J Reynolds, 'Empathy and quality of care.', *The British Journal of General Practice*, **52**(Suppl), S9, (2002).

[9] Katherine Miller, Marty Birkholt, Craig Scott, and Christina Stage, 'Empathy and burnout in human service work an extension of a communication model', *Communication Research*, **22**(2), 123–147, (1995).

[10] Jayne Morgan and Kathleen Krone, 'Bending the rules of" professional" display: Emotional improvisation in caregiver performances', *Journal of Applied Communication Research*, **29**(4), 317–340, (2001).

[11] Bilge Mutlu and Jodi Forlizzi, 'Robots in organizations: The role of workflow, social, and environmental factors in human-robot interaction', in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, HRI '08, pp. 287–294, New York, NY, USA, (2008). ACM.

[12] Stanley Schachter, 'The interaction of cognitive and physiological determinants of emotional state', *Advances in experimental social psychology*, **1**, 49–80, (1964).

[13] Amanda Sharkey and Noel Sharkey, 'Granny and the robots: ethical issues in robot care for the elderly', *Ethics and Information Technology*, **14**(1), 27–40, (2012).
[14] Allen C Smith III and Sherryl Kleinman, 'Managing emotions in medical school: Students' contacts with the living and the dead', *Social Psychology Quarterly*, 56–69, (1989).
[15] Robert Sparrow and Linda Sparrow, 'In the hands of machines? the future of aged care', *Minds and Machines*, **16**(2), 141–161, (2006).
[16] Noreen M Sugrue, 'Emotions as property and context for negotiation', *Journal of Contemporary Ethnography*, **11**(3), 280–292, (1982).
[17] Sherry Turkle, Will Taggart, Cory D Kidd, and Olivia Dasté, 'Relational artifacts with children and elders: the complexities of cybercompanionship', *Connection Science*, **18**(4), 347–361, (2006).
[18] Kazuyoshi Wada and Takanori Shibata, 'Living with seal robotsits sociopsychological and physiological influences on the elderly at a care house', *Robotics, IEEE Transactions on*, **23**(5), 972–980, (2007).
[19] Wendell Wallach and Colin Allen, *Moral machines: Teaching robots right from wrong*, Oxford University Press, 2008.
[20] A. Weiss and T. Lorenz, 'Icsr 2013 workshop 3: Final report and results: Taking care of each other: Synchronization and reciprocity for social companion robots', Technical report, Workshop report available at: http://workshops.acin.tuwien.ac.at/ISCR2013/, (2013).
[21] Robert B Zajonc, 'Feeling and thinking: Preferences need no inferences.', *American psychologist*, **35**(2), 151, (1980).
[22] Slavoj. Zizek, Elizabeth (eva Elizabeth) Wright, and Edmond Leo Wright, *The Zizek Reader*, Blackwell, 1999.