(Un)learning Reflection with the Help of Moral Agents

Potential risks on an individual and societal level to be considered in interaction design

Lara Christoforakos*

Ludwig-Maximilians-Universität München, Department of Psychology, lara.christoforakos@psy.lmu.de

Sarah Diefenbach

Ludwig-Maximilians-Universität München, Department of Psychology, sarah.diefenbach@psy.lmu.de

Daniel Ullrich

Ludwig-Maximilians-Universität München, Department of Computer Science, daniel.ullrich@ifi.lmu.de

Interactive technologies are increasingly applied to initiate human behavior change towards socially aspired outcomes. Within this, the sub-category of so-called moral agents are proposed as counterpart technologies that go beyond supporting the user to achieve personally set goals and actively intervene based on inscribed values. On the one hand, the concept of moral counterparts appears promising, especially within domains where specific behavioral goals might be rather abstract or uncomfortable. On the other hand, it can come with new challenges which could diminish individual reflection and sustainable behavior change in the long run. Our position paper discusses such challenges on individual and societal level as well as first design ideas that could play a role in this, to be explored in the workshop and beyond.

CCS CONCEPTS • Human-centered computing • Human computer interaction (HCI) • HCI theory, concepts and models

Keywords: moral agents, psychological effects, counterpart technologies, autonomy, responsibility, reflection

ACM Reference Format:

First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1 MORAL AGENTS AS A NEW TYPE OF COUNTERPART TECHNOLOGIES

Interactive technologies are increasingly applied to initiate human behavior change towards socially aspired outcomes. Smartwatches support us in becoming healthier by tracking our daily calorie intake and movement. Various apps help us become more efficient in time management. One can even find technologies that measure the amount of water spent while showering and support the user in behaving more sustainably.

---

* Place the footnote text for the author (if applicable) here.

In the context of Human-Computer Interaction (HCI), such technologies have mostly been known under the term Persuasive Technologies [3]. These technologies support users in achieving their individually set goals. They do so as they are developed based on, amongst others, theoretical concepts of behavior change that, for example, imply methods of goal setting [7]. In the domain of sustainability, specific examples are technologies that confront the user with information on how much (hot) water is consumed during a shower [5] or visualize water as a limited resource that decreases [6]. In these cases, the user might set an individual limit for the water spent while showering and there are no direct consequences that originate from the technology if this limit is exceeded.

The concept of moral agents as "more full-fledged social actors," as proposed in this workshop, addresses a new type of technologies that, first, actively engage in interactions and act according to their inscribed values. Second, they involve humans in a moral dialogue regarding their behavior, and third, they express moral demands on them, which can originate from the technology itself or other entities. They do not represent simple tools but act as counterparts with whom the user can interact or cooperate with. Transferring this concept to the example of saving water, an according technology might advise the user not to exceed a certain amount of water spent, according to general environmental calculations. Moreover, it might even actively intervene through alarming sounds, visualize potential environmental consequences via screen and even turn off the water after a certain amount spent. Thus, the technology would start a sort of moral dialogue with the user.

Such qualities of moral agents might lead to an increase in the degree of autonomy users attribute to these technologies. According to the psychological self-determination theory, perceived autonomy, i.e., the "inner endorsement of one's actions, the sense that they emanate from oneself and are one's own" [2], can affect the responsibility people attribute to their counterpart for a behavioral outcome. This principle also seems to apply to HCI. Studies have shown that participants attributed more responsibility to a computer that behaved autonomously (by providing real-time advice through an interface agent) compared to a computer that behaved non-autonomously (by providing a help menu, [9]). Moreover, based on previous research, a robot's perceived agency (measured by questions about the robot's control over the situation and its ability to make its own decisions) can affect the responsibility attributed to the robot for its actions [10]. Considering moral agents, this means that the more autonomous these technologies are perceived due to their active choices and initiation of interaction, the more responsible users might find the technology for the outcome of their interaction with such. Focusing on the context of sustainability, this can form challenges on an individual and societal level.

2 RISKS ON INDIVIDUAL AND SOCIETAL LEVEL

From a psychological perspective it appears particularly worthwhile to explore and discuss risks of moral agents in enabling users to behave sustainably in a self-dependent manner and based on their ability to reflect. In the following, potential risks are discussed on an individual and societal level.

2.1 The lazy individual

On an individual level, based on the phenomenon of diffusion of responsibility, the feeling of responsibility for an outcome can decrease with an increasing number of people involved in a social situation [11]. Assuming a certain transferability of social phenomena to HCI (e.g., [8]), when a moral agent cooperates with a user in performing a certain behavior, for example, acting more sustainably, the user might feel less responsible for the behavioral outcome. In turn, the user might become less involved and sort of lazy with regard to caring for sustainability in the frame of the considered activity. For example, a moral agent might support users in consuming less water while showering, by reminding them to turn off the water or even actively turning it off in between. After a while, the user might get used to the technology actively intervening

and therefore become less self-conscious regarding the amounts of water consumed and perceive less responsibility for the outcome of their actions. The effect of diffusion of responsibility could be even more potent in the context of moral agents because users could presumably perceive such an agent not as merely one social actor but as a representation of many. This assertion is based on the fact that moral agents would be a result of the effort put in by many people: designers, coders, developers as well as a wide bunch of scientists which worked on the models leading to the embedded moral impetus. The result would be even less perceived responsibility from the user's point of view.

At the same time, there could be negative consequences for the user's self-efficacy [1] and, in turn, the general motivation to behave sustainably. That is, when the user shows sustainable behavior as a result of interaction with a moral agent, they might not perceive this outcome to be their own merit, which might not be a great motivation to continue showing this behavior.

2.2 The obedient society

On a broader, societal level, if moral agents that, for example, support users in showing sustainable behavior, become the norm, people might generally get used to counterpart technologies intervening in situations, where people have the capacity to behave more sustainably and therefore become less sensible or even blind regarding such situations themselves. Such an outcome might even come with the challenge of a possible general reduction of individuals' reflection regarding their own actions and potential effects on society.

Developing the thought further, if members of a society get used to the existence of moral agents, a sort of hierarchy might slowly become established. There, individuals might tend to heed the advice of moral agents and, with time, become rather obedient and stop questioning the advice's quality and rightness. At the same time, users might put less effort into developing and refining their own opinions and attitudes regarding important societal issues such as sustainability.

If moral agents became more and more spread in society, individuals could even adapt a habit of internalized obedience which could appear as some sort of semi-automated behavior, for example, if individuals spot a known moral agent in an unknown place like a restaurant's restroom, they might automatically show the requested behavior. Conversely, based on a lack of internalized moral standards, in the absence of such agents, the requested behavior would fail to appear.

3 OUTLOOK

Overall, it appears promising to support users in behavior change regarding societally essential domains, such as sustainability, through technologies. The above-mentioned risks regarding users' enablement to reflect and self-dependently behave sustainability, might appear overdramatized at first. Still, it is important to focus on such downsides of the interaction with this new type of counterpart technologies. Furthermore, it seems essential to explore and in turn develop potential design solutions to counteract such risks and use the full potential of moral agents to empower users to behave sustainably throughout different domains.

A first thought regarding the diffusion of responsibility through the cooperation with moral agents could be to focus on the extent and means of designing moral agents as "social" counterparts. Previous research has, for example, shown anthropomorphic products to be attributed more responsibility when compared to non-anthropomorphic products (e.g., [4]). Therefore, the questions for future research emerge: What difference does it make if the moral agent intervening in unsustainable shower behavior uses the medium of voice vs. just a screen to intervene? What role does the perception of an "own will" or inscribed values play? Do users potentially become even lazier the more they attribute an "own will" or values to the technology as they know the agent will "worry" about the observed water consumption? Is there a sweet spot to designing this type of agency?

ACKNOWLEDGMENTS

Part of this research was funded by the German Federal Ministry for Education and Research (BMBF), Project MOVEN (FKZ: 01UU2204B) as well as the German Research Foundation (DFG), Project PerforM (425412993) as part of the Priority Program SPP2199 Scalable Interaction Paradigms for Pervasive Computing Environments.

REFERENCES

[1] Bandura Albert. 1977. Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review, 84(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

[2] Edward L. Deci and Richard M. Ryan. 1987. The support of autonomy and the control of behavior. Journal of Personality and Social Psychology, 53, 1024–1037. https://doi.org/10.1037/0022-3514.53.6.1024

[3] Brian J. Fogg. 2002. Persuasive technology: using computers to change what we think and do. Ubiquity, (December 2002), 2. https://doi.org/10.1145/764008.763957

[4] Pamela J. Hinds, Teresa L. Roberts, and Hank Jones. 2004. Whose job is it anyway? A study of human-robot interaction in a collaborative task. Human-Computer Interaction, 19, 151–181. https://doi.org/10.1207/s15327051hci1901&2_7

[5] Karin Kappel and Thomas Grechenig. 2009. "show-me": water consumption at a glance to promote water conservation in the shower. In Proceedings of the 4th International Conference on Persuasive Technology - Persuasive '09. ACM, New York, 1-6. https://doi.org/10.1145/1541948.1541984

[6] Matthias Laschke, Marc Hassenzahl, Sarah Diefenbach, and Marius Tippkämper. 2011. With a little help from a friend: a shower calendar to save water. In Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA. ACM, New York, 633-646. https://doi.org/10.1145/1979742.1979659

[7] Edwin A. Locke and Gary P. Latham. 1990. A theory of goal setting & task performance. Prentice-Hall, Inc. New Jersey, USA.

[8] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In B. Adelson, S. Dumais, & J. Olson (Eds), CHI '94 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, New York, 72–78. https://doi.org/10.1145/191666.191703

[9] Alexander Serenko. 2007. Are interface agents scapegoats? Attributions of responsibility in human-agent interaction. Interacting with Computers, 19, 293–303. https://doi.org/10.1016/j.intcom.2006.07.005

[10] Sophie van der Woerdt and Pim Haselager. 2019. When robots appear to have a mind: The human perception of machine agency and responsibility. New Ideas in Psychology, 54, 93–100. https://doi.org/10.1016/j.newideapsych.2017.11.001

[11] Lioba Werth and Jennifer Mayer. 2008. Sozialpsychologie. Springer. Heidelberg, Germany.