Janina Loh, Wulf Loh (eds.)

SOCIAL ROBOTICS AND THE GOOD LIFE

The Normative Side of Forming Emotional Bonds With Robots

[transcript] Philosophy

Janina Loh, Wulf Loh (eds.)
Social Robotics and the Good Life

Philosophy

Janina Loh (née Sombetzki) is an ethicist (Stabsstelle Ethik) at Stiftung Liebenau in Meckenbeuren on Lake Constance. They have been a university assistant (PostDoc) at Christian-Albrechts-Universität Kiel (2013-2016) and in the field of philosophy of technology and media at Universität Wien (2016-2021). Their main research interests lie in the field of critical posthumanism, robot ethics, feminist philosophy of technology, responsibility research, Hannah Arendt, theories of judgement, polyamory and poly-ethics as well as ethics in the sciences.

Wulf Loh is an assistant professor at the Int. Center for Ethics in the Sciences and Humanities (IZEW) at Eberhard Karls Universität Tübingen and supervisor of various technology development projects. His areas of expertise are within applied ethics, social philosophy, political philosophy, and philosophy of law.

Janina Loh, Wulf Loh (eds.)

Social Robotics and the Good Life

The Normative Side of Forming Emotional Bonds With Robots

[transcript]

The free availability of the e-book edition of this publication was financed by the Fachinformationsdienst Philosophie.

FACHINFORMATIONSDIENST
PHILOSOPHIE

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de

First published in 2023 by transcript Verlag, Bielefeld
© Janina Loh, Wulf Loh (eds.)

Cover layout: Kordula Röckenhaus, Bielefeld
Printed by Majuskel Medienproduktion GmbH, Wetzlar
Print-ISBN 978-3-8376-6265-8
PDF-ISBN 978-3-8394-6265-2
https://doi.org/10.14361/9783839462652
ISSN of series: 2702-900X
eISSN of series: 2702-9018

Printed on permanent acid-free text paper.

Contents

Introduction – Social Robotics and the Good Life
The Normative Side of Forming Emotional Bonds With Robots
Janina Loh, Wulf Loh 7

Part I – Understanding, Defining, Conceptualizing

Virtues, Robots, and Good Lives: Who Cares?
Charles M. Ess 25

The Relational Turn
Thinking Robots Otherwise
David J. Gunkel 55

From Tool Use to Social Interactions
Anna Strasser 77

Reflecting (on) Replika
Can We Have a Good Affective Relationship With a Social Chatbot?
Eva Weber-Guskar 103

Part II – Design, Imitation, Trust

You Can Love a Robot, But Should You Fight With it?
Perspectives on Expectation, Trust, and the Usefulness of Frustration Expression in Human-Machine Interaction
Jacqueline Bellon, Tom Poljanšek 129

Empathic Machines?
Ethical Challenges of Affective Computing from a Sustainable Development Perspective
Cordula Brand, Leonie N. Bossert, Thomas Potthast 157

Part III – Care, Love, Sex

Granny and the Sexbots
An Ethical Appraisal of the Use of Sexbots in Residential Care Institutions for Elderly People
Karen Lancaster 181

Alice Does not Care
Or: Why it Matters That Robots "Don't Give a Damn"
Imke von Maur 209

Emotional Embodiment in Humanoid Sex and Love Robots
Cindy Friedman, Sven Nyholm, Lily Frank 233

Contributors 257

Introduction – Social Robotics and the Good Life
The Normative Side of Forming Emotional Bonds With Robots

Janina Loh, Wulf Loh

Robots and Their Tasks in Society

Robots have existed as a concept that is still common today for a hundred years. Historically, the term "robot" goes back to the Czech word "robota", which stands for work, hard labor, and forced labor. Karel Čapek first used it in the play R.U.R. Rossum's Universal Robots (1921) to refer to humanoid contraptions that are at the service of humans. This historical vision of the robot as an artificial slave subsequently was the guiding idea shaping the development of robotics.[1]

Without doubt, robots can now be found in many areas of human life. Industry was the first sector they made their entrance into around the middle of the 20th century with the *Unimate*[2] robot. Particularly in this area, robots are assigned to jobs that are considered "dull, dangerous, and dirty". Although it is anything but clear which activities fall under this category (Marr 2017), any socio-ethical doubts arising from this have not been able to put a stop to the advancing robotization of industrial production. Today, tasks that are considered boring because they are repetitive and monotonous, as well as dirty

[1] By "robot", we mean an electro-mechanical machine that a) has some form of independent body, b) possesses at least one processor, c) is equipped with sensors that collect information about the world d) as well as with at least one effector that translates signals into mechanical processes. A robot's behavior e) is or at least appears autonomous, enabling it to f) interact with or influence its environment (cf. Loh 2019a: 7; Misselhorn 2013: 43).

[2] A composite of "universal" and "automation"; constructed by Joseph Engelberger in 1961.

and even dangerous, are increasingly being performed by robots on assembly lines, in production halls and warehouses around the globe. The Kuka robots in the automotive industry and the Amazon warehouse robots are but some recent examples of this phenomenon.

Social Robotics and Robots as Social Companions

Robots are increasingly entering the personal sphere of everyday human interaction. This area of so-called *social robotics* is characterized by a large number of everyday activities, that at the same time can be very private and sensitive for various reasons (Breazeal 2002; Duffy 2008, 2004; Fong et al. 2003; Markowitz 2015; Seibt et al. 2016; Fronemann et al. 2022). The main fields in which robots are employed here are therapy and care, specifically in the form of activation and keeping company, but also education and even sexuality, friendship, and love. In those areas, the attribution of the three Ds no longer applies. Even when the robots are primarily designed as auxiliary tools for daily tasks, the typical activities are marked by interdependence and physical and emotional closeness.

From robots for sales assistance such as Paul (who guides customers through the aisles of the electronics retailer Saturn), care robots like Care-O-Bot, entertainment robots such as Pepper, to sex robots (Markowitz 2015: 41), are all examples for the growing group of social robots that are used in close proximity to humans. Depending on their respective tasks and the extent to which they enter into direct interaction with humans, they need to possess social skills in some form. This anthology is dedicated to them and to the question about the relationships they allegedly form or are supposed to be able to form with humans.

In this regard, the authors of this anthology take a closer look at three topics of social robotics in particular. *The first part* revolves around questions of defining and understanding basic elements of human-robot interaction (HRI). Charles Ess, David Gunkel, Anna Strasser, and Eva Weber-Guskar examine the basic anthropological and ontological assumptions of HRI, ask what we mean by “social agent” and, correspondingly, by “artificial social agent”, address whether interaction with robots should be reconsidered in general, and when it is morally appropriate to speak of successful HRI.

The second part deals with questions of the design of robots as social companions, the imitation of emotions, and corresponding or associated reactions such as trust and expectations with regard to the behavior of a robot. Cordula Brand, Leonie N. Bossert, Thomas Potthast, Jacqueline Bellon, and Tom Poljanšek in their contributions cover the ethical implications that arise from the embodied and anthropomorphic design of robots.

Finally, in the *third part*, the authors address whether specific emotional relationships with robots are possible, namely caring, loving, and sexual relationships. Imke von Maur, Lily Frank, Cindy Friedman, Sven Nyholm, and Karen Lancaster are concerned with the possibility and ethical appropriateness of these forms of relationships and its repercussions.

Part I – Understanding, Defining, Conceptualizing: Robots as Social Companions

Already Čapek raises numerous philosophical, ethical, and anthropological questions in his "R.U.R. Rossum's Universal Robots". These include, for example, the nature of man, the responsibility of scientists, and what it means to form an emotional bond with another being. For instance, his piece ends with the prospect of a love affair that seems to be developing between two robots. Thus, in the historical understanding of the robot established by Čapek, a broad foundation is already laid for the discussions that were to arise in the decades that followed and that extend into social robotics. Is it possible to live a good, successful life with and around robots, including forming intimate romantic and sexual relationships with them?

In his text "Virtues, Robots, and Good Lives: Who Cares?", *Charles Ess* deals with these questions, which also concerned Čapek, from a virtue ethics perspective. Maybe a reformulated, "relational" virtue ethics, supplemented by an "ethical pluralism", can help circumvent the problems of an "ethical relativism and computer-mediated colonization"? By way of applying virtue ethics to sex robotics as an example of emotionally intimate relationships with robots, Ess is skeptical, however, that a sexual human-machine relationship can ever be said to be truly "complete." On the other hand, it is too quick to simply depreciate robots on the basis of, say, their lack of autonomy, as this ultimately leads to a "reinscribing of traditional patriarchal and racist attitudes". Therefore, it is necessary to further develop a "pluralistic" ethics of virtue into an "ethics of care" appropriate to the current developments in social robotics.

In order to answer the question whether relationships with robots are desirable, David Gunkel, Eva Weber-Guskar and Anna Strasser, first address the status of robots in general. Can we understand them as agents with whom genuine relationships are possible or are robots unable to transcend their traditional object status? With these considerations, the authors tie in with a broad tradition in robot ethics, in which basically three currents can be distinguished.

Within the first research area, authors ask to what extent robots can be considered as potential moral agents (Floridi/Sanders 2004; Misselhorn 2013; Moor 2006; Sullins 2006; Wallach/Allen 2009). Accordingly, they consider the degree to which robots are capable of moral action and which competencies they must possess to this end. Depending on the respective understanding of agency, morality, and the competencies to be realized for this purpose, this includes the attribution of freedom and autonomy as a condition for moral action, cognitive competencies (such as thinking, mind, reason, judgment, intelligence, consciousness, perception, and communication), but also empathy and emotions. Defining the "minimal conditions" for understanding robots as "social agents" is also *Anna Strasser's* concern in her text "From Tool Use to Social Interactions". For this purpose, she considers the case of "joint actions" as an example for "social interaction" and seeks to establish a more appropriate understanding of HRI or the relationships we enter into with robots.

Within the second area, authors deal with the question whether robots should be regarded as moral patients, i.e. as objects of moral consideration (Damiano/Dumouchel 2018; Darling 2012, 2017; Duffy 2003, 2004, 2008; Gerdes 2017; Johnson 2011; Tavani 2018). These approaches are concerned with how to deal with artificial systems, what kind of moral value they may have, even if they may be incapable of moral agency themselves. Topics include, for example, the formulation of codes of ethics in corporations, the desirability and possibility of relationships with and to robots, the question of exploiting or "enslaving" robots, or the assessment of the use of robots for therapeutic purposes. Some thinkers discuss the possibility of ascribing rudimentary rights to some types of robots. Just as Immanuel Kant in § 17 of the second part of his Metaphysics of Morals is opposed to cruelty towards animals, because this would lead to morally questionable attitudes in us humans, Kate Darling, for example, argues in favor of robot rights because people are then more likely to maintain their moral virtues in other interactions as well.

Nonetheless, the fact that robots are regarded merely as moral patients does not preclude the possibility of humans having emotional relationships with them. Against the background of this possibility, *Eva Weber-Guskar* in her text “Reflecting (on) Replika: Can We Have a Good Affective Relationship With a Social Chatbot?” rejects the possibility of a relationship comparable to that between humans, using the example of the social chatbot Replika. However, Weber-Guskar also emphasizes that “the lack of emotional mutuality” does not justify a general rejection of the possibility of affective relationships with “social or emotional(ized) AI”.

A third strand within robotics ethics transcends the obvious dichotomy between moral agency and patiency. Authors here discuss alternatives to the classical distinction between subjects and objects of moral action. Within the framework of these “inclusive” or “inclusivist” approaches (Loh 2019a, 2020, 2022), the focus is on problems with a traditional conception of the (human) person that underlies the notion of the moral agent. The understanding of the human being as the core of ethical thinking, as the main moral agent, as the pivot of the attribution of abilities, competences, and values is questioned and challenged in these inclusive approaches.

In his text “The Relational Turn: Thinking Robots Otherwise,” *David Gunkel* describes the project of such an inclusive approach in order to “introduce and formulate a meta-ethical theory”. First, he takes a look at the peculiarities of the classical, exclusive ethical positions. In a second step, he then outlines his alternative of a relational Thinking Otherwise, and in a third step meets possible objections. Gunkel is thus concerned with a general new understanding of the possibility of entering into relationships with robots. He shows that questions of moral status of and potential relationships with robots have less to do with the robots themselves and more to do with “us and the limits of who is included in and what comes to be excluded from that first-person plural pronoun, ‘we.’”

Part II – Design, Imitation, Trust: Anthropomorphization and its Function for Social Robotics

Robots as social companions can only carry out their activities in close proximity to humans, if they are accepted by people in their immediate private space. People do not want to be assisted in their personal hygiene by scary or repulsive machines, they do not want to be cared for and touched mentally and physically by cold apparatuses – this seems a pretty straightforward assumption and is therefore at the bottom of almost all social robot design. The latter is quasi unanimously catered to an image of robots that their human users can identify with and thus engage with more easily. They are, according to Kate Darling, “specifically designed to socially interact with humans” (2012). Such a trust-inspiring design is in many cases anthropomorphic (i.e., human-like), or more rarely, zoomorphic (i.e., animal-like).

The anthropomorphization of non-human entities does not only concern their outer form, but can also refer to their behavior and thus to the attribution of human competencies (Fink 2012: 200). Therefore, in the field of social robotics, a distinction is sometimes made between anthropomorphic design, which primarily comprises externally perceptible criteria such as “shape, speech capabilities, facial expression,” and the like, and anthropomorphic interaction design, which targets the “social phenomenon that emerges from the interaction between a robot and an [sic!] user” and is sometimes called “anthropomorphism” (in the proper sense) (all citations in Lemaignan et al. 2014: 226; cf. Złotowski et al. 2015).

Our capacity for anthropomorphization seems to be a psychological fact and thus primarily a topic of the social sciences, psychology, and (in the context of robots) science and technology studies. However, the anthropomorphic lens through which we often view and evaluate the nonhuman world frequently serves as a vehicle for placing humanized beings in the moral universe (see *Eva Weber-Guskar's* text in this volume for more on this). After all, the more human we assess a counterpart, the more similar we make it to ourselves, the more we identify with it, the more willing we are also to assign it a (moral) value similar to that of humans. Indeed, we are forced to do so to a certain extent if we do not want to be argumentatively inconsistent. The anthropomorphic gaze gives a moral value to all beings it encounters.

In psychology, anthropomorphization is traditionally viewed in a negative light, “as a bias, a category mistake, an obstacle to the advancement of knowledge, and as a psychological disposition typical of those who are immature and unenlightened, i.e., young children and ‘primitive people’” (Damiano/Dumouchel 2018: 2; cf. Duffy 2003: 180-181). Some authors even go a step further by declaring humanized robots to be a kind of “‘cheating’ technology” that is ethically problematic (Turkle 2011: 514). Humanized robots not only deceive us into believing that they possess mental states, but also into the “illusion of relationship” that we can actually only enter into with humans (Turkle 2005: 62; cf. Damiano/Dumouchel 2018: 1; Lin 2012: 11).

This argument of a deceptive technology or a “culture of simulation” (Turkle 2011) can be interpreted in terms of virtue ethics, insofar as the good life importantly also depends on human relationships (Nussbaum 2007). In this sense, humans commit a moral error when they replace human-human relationships as the genuine form of relationship with a mirage evoked by a robot (see also the text by *Charles Ess* in this volume).[3] Accordingly, a "simulated feeling is never feeling, simulated love is never love" (Turkle 2010: 4; cf. Damiano/Dumouchel 2018: 5).

The traditionally negative connotation of anthropomorphism in psychology[4] is countered by studies in social robotics that contrast these concerns with a positive interpretation of the human capacity for anthropomorphization. For example, Luisa Damiano and Paul Dumouchel outline an optimistic approach that views anthropomorphism not as a "cognitive error" but "as a fundamental tool" (2018: 5) that can support and enhance HRI. They state that what makes social robots special is that they "tend to blur the traditional ontological categories that humans use to describe the world," most notably the subject-object dichotomy, but also the categories of animate and inanimate, sentient and non-sentient (2018: 4) (on overcoming the subject-object dichotomy in inclusive robot ethics approaches, see also the text by *David Gunkel* in this volume).

Other authors view anthropomorphizing robots only as "desirable where it enhances the function of the technology" (Darling 2017: 174). Humans can and should identify with social robots used in elder care and households, as these machines interact with their owners in an intimate way. Otherwise, they would be unable or unwilling to engage with the artificial system. On the other hand, authors conclude that a robot must "not be too similar to a human being if it is supposed to elicit empathy" (Misselhorn 2009: 117). Often referring to the so-called "uncanny valley" (Mori 1970: 33-35), they argue that otherwise, the inability to clearly categorize the robot adequately according to such important categories as animate/ inanimate will irritate, repel, and scare us.

It is interesting to note, however, that with regard to the question of whether it makes sense to design robots in such a way that people can engage

[3] The argument against anthropomorphism can also be spelled out deontologically and utilitarian: deontologically, if I neglect my duties towards other people, for example, because of my emotional attachment to a robot; and utilitarian, if I should, for example, attribute the ability to suffer to robots due to anthropomorphizing them, and therefore include them in the calculation of total utility.

[4] There is also a positive interpretation of anthropomorphism in psychology, pointing out that under the anthropomorphic gaze nonhuman beings can become "familiar, explainable, or predictable" (Fink 2012: 200).

and form relationships with them, the focus is primarily on positive emotional forms of relationships such as care, love, and friendship. As discussed earlier, negative affective relationships are mostly used as reasons for rejecting an anthropomorphic design. In their text "You Can Love a Robot, But Should You Fight With it?", *Jacqueline Bellon* and *Tom Poljanšek* raise the question of whether such "frustrated-related concepts of human emotion" can also be at times conducive to a good life. If this is the case, the simulation of negative emotions should not be excluded in HRI.

In social robotics, the question of appropriate design of robots has so far been limited to direct interaction in close proximity between humans and machines. In their text "Empathic Machines? Ethical Challenges of Affective Computing from a Sustainable Development Perspective", *Cordula Brand, Leonie N. Bossart* and *Thomas Potthast* extend this narrow perspective using a justice-based approach, namely the "Sustainable Development framework". By doing so, they evaluate whether and how the presumed advantage of "affective computing" can actually be realized for all humans in a just and needs-based way.

Part III – Care, Love, and Sex With Robots as Social Companions

Depending on the status we are willing to ascribe to robots (Part I of this volume) and depending on their design, which can be attractive or repulsive to us (Part II of this volume), some people actually enter into friendly, sexual, caring, or loving relationships with robots. In a very literal sense, it is sex robots that are closest to us.

For most people, sex robotics may sound like pure science fiction. In reality, however, there are already several large international companies that mass-produce and sell sex robots – including two Chinese companies (DS Doll Robotics and Shenshen All Intelligent Technology Co.) and one from the USA (RealDoll with RealBotix). TrueCompany in 2010 was the first company in the world to launch a sex robot called Roxxxy. Roxxxy had interactive capabilities such as, according to the now offline homepage, "hear what you say, speak, feel your touch, move their bodies, are mobile and have emotions and a personality." Roxxxy was said to be able to develop its own personality (or as many different roles as desired) through interaction with its users. But it was also possible to give her one of five pre-programmed personalities. Besides that, she could be given different hairstyles and hair colors. Aside from the aforementioned skills, Roxxxy should have also been able to "listen, talk, carry on a conversation and feel your touch" and even "have an orgasm" (TrueCompanion 2019).

Other examples include Shenshen all Intelligent Technology's Emma robot, Matt McMullen's sex robot Harmony (Realbotix), Samantha that is supposedly equipped with a "moral code", as well as LumiDoll's sex robot Kylie (Mlot 2018; Morgan 2017). It is clear that in sex robots highly questionable gender stereotypes are oftentimes upheld and heteronormative, patriarchal, instrumentalizing, and discriminatory power structures are confirmed (see also the text by *Charles Ess* in this volume). As a result, the spectrum of ethical issues is evident (Cheok et al. 2017; Danaher 2017; Danaher/McArthur 2017; Kubes 2019; Levy 2012, 2008; Loh 2019b; Scheutz 2012; Whitby 2012).

In her text "Granny and the Sexbots", *Karen Lancaster* deals with the possibility and moral desirability of using sex robots in elder care. In doing so, she addresses not one but two taboo topics – aside from sex robotics, the fact that older people may also have a wish to experience a fulfilled sexuality.

At the same time, Lancaster's text bridges the gap between two fields of social robotics, namely between sex robotics on the one hand and medical, therapeutic and nursing robotics on the other. Here, too, the use of sex robots is discussed. But apart from sex robots, numerous assistance systems are already being used today to support caregivers in the medical, therapy, and nursing sectors in their often extremely physically and mentally demanding work. From lifting and transport systems to companion robots that activate, entertain, and thereby reduce loneliness, to therapy robots that promote communication with patients, a broad spectrum of artificial systems exists for a variety of different tasks in hospitals, therapy, and care facilities. One field that has been developing steadily for a good 15 years, for example, is the robot use in the therapy of children with autism (Richardson et al. 2018; critically Elder 2017).

The artificial seal Paro is a good example of a zoomorphically designed care assistance robot. Paro is modeled after a young harp seal and mainly used in geriatric care and therapy, especially for people with severe dementia. These people particularly tend to isolate themselves from their human caregivers, but often open up to animals. As a robot, Paro cannot be accidentally hurt as it might happen to real animals. Nonetheless, it is said to offer many of the benefits of a regular human-animal-interaction therapy in these application contexts (Shibata/Wada 2011; Wada et al. 2008).

Because of the numerous challenges associated with robots as potential social companions, some thinkers are skeptical about the question of whether it makes sense or is morally desirable to develop robots in such a way that humans want to form relationships with them. In this vein, in her text "Alice Does not Care. Or: Why it Matters That Robots 'Don't Give a Damn', *Imke von Maur* also rejects the use of care robots with the aim "to reduce loneliness". Since "real care" involves both meanings of the word, robots are according to her not capable of caring in this sense. Even more, people run the risk of "giving up expectations of real care and true relationships" when they get involved with robots.

The debates about the pros and cons of the possibility and especially the moral desirability (in terms of a good life) of emotional relationships with robots are pointedly summarized in the text "Emotional Embodiment in Humanoid Sex and Love Robots" by *Cindy Friedman, Sven Nyholm,* and *Lily Frank*. By way of three ethical questions, they discuss some of the central approaches to robot ethics, and in doing so bring to the forefront the challenges we face in the theoretical conception (cf. Part I of this volume), the actual design (cf. Part II), and the practical use of robots as potential social companions (cf. Part III).

The authors of this anthology show that the challenges we face regarding the questions of the good life and the possibility of emotional relationships with robots are undoubtedly manifold. But they also show that addressing these questions is worthwhile as we become more sensitive to the complexity of a society in the age of digitalization, automation, and robotization.

References

Ackermann, Evan (2015): "Care-O-bot 4 Is the Robot Servant We All Want but Probably Can't Afford«, in: IEEE Spectrum; https://spectrum.ieee.org/automaton/robotics/home-robots/care-o-bot-4-mobile-manipulator.

Anderson, Michael/Anderson, Susan Leigh/Armen, Chris (2006a): "MedEthEx. A Prototype Medical Ethics Advisor." In: Proceedings of the Eighteenth Conference on Innovative Applications of Artificial Intelligence, Boston, Massachusetts, pp. 1759-1765.

Anderson, Michael/Anderson, Susan Leigh/Armen, Chris (2006b): "An Approach to Computing Ethics", in: Intelligent Systems IEEE 4, pp. 2-9.

Breazeal, Cynthia (2002): Designing Sociable Robots, Cambridge MA, London: The MIT Press.

Cheok, Adrian/Karunanayaka, Kasun/Yann Zhang, Emma (2017): "Lovotics. Human-Robot Love and Sex Relationships." In: Patrick Lin/Ryan Jenkins/Keith Abney (eds.), Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence, New York: Oxford University Press, pp. 193-213.

Damiano, Luisa/Dumouchel, Paul (2018): "Anthropomorphism in Human-Robot Co-evolution." In: Frontiers in Psychology 9, pp. 1-9; https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00468/full#B54.

Danaher, John (2019): "Should We Be Thinking about Robot Sex?" In: John Danaher/Neil MaArthur (eds.), Robot Sex. Social and Ethical Implications, Cambridge, Massachusetts, London: The MIT Press, pp. 3-14.

Danaher, John/McArthur, Neil (eds.) (2017): Robot Sex. Social and Ethical Implications, Cambridge, Massachusetts, London: The MIT Press.

Darling, Kate (2012): "Extending Legal Protection to Social Robots." In: IEEE Spectrum, September 10 2012; https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/extending-legal-protection-to-social-robots.

Darling, Kate (2017): "'Who's Jonny?' Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy." In: Patrick Lin/Ryan Jenkins/Keith Abney (eds.), Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence, New York: Oxford University Press, pp. 173-188.

Duffy, Brian R. (2003): "Anthropomorphism and the social robot." In: Robotics and Autonomous Systems 42, pp. 177-190.

Duffy, Brian R. (2004): "Social Embodiment in Autonomous Mobile Robotics." In: International Journal of Advanced Robotic Systems 1, pp. 155-170.

Duffy, Brian R. (2008): "Fundamental Issues In Affective Intelligent Social Machines." In: Open Artificial Intelligence Journal 2, pp. 21-34.

Elder, Alexis (2017): "Robot Friends for Autistic Children. Monopoly Money." In: Patrick Lin/Ryan Jenkins/Keith Abney (eds.), Robot Ethics 2.0. From Autonomous Cars to Artificial Intelligence, New York: Oxford University Press, pp. 113-126.

Fink, Julia (2012): "Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction." In: Shuzhi Sam Ge/Oussama Khatib/John-John Cabibihan/Reid Simmons/Mary-Anne Williams (eds.), Social Robotics. 4th International Conference, ICSR 2012, Chengdu, China, October 2012. Proceedings, Berlin, Heidelberg: Springer, pp. 199-208.

Floridi, Luciano/Sanders, J. W. (2004): "On the Morality of Artificial Agents." In: Minds and Machines 14, pp. 349-379.

Fong, Nourbakhsh/Terrence, Illah/Dautenhahn, Kerstin (2003): "A survey of socially interactive robots." In: Robotics and Autonomous Systems 42, pp. 143-166.

Fronemann, Nora/Pollmann, Kathrin/Loh, Wulf (2022): "Should my robot know what's best for me? Human–robot interaction between user experience and ethical design", in: AI & Society 37, pp. 517-533.

Gerdes, Anne (2017): "The Issue of Moral Consideration in Robot Ethics." In: ACM SIGCAS Computers & Society 45, pp. 247-279.

Johnson, Deborah G. (2011): "Computer Systems. Moral Entities but Not Moral Agents." In: Michael Anderson/Susan Leigh Anderson (eds.), Machine Ethics, New York, pp. 168-183.

Kubes, Tanja (2019): "Bypassing the Uncanny Valley. Sex Robots and Robot Sex Beyond Mimicry." In: Janina Loh/Mark Coeckelbergh (eds.), Feminist Philosophy of Technology, Stuttgart: J.B. Metzler, pp. 59-73.

Lemaignan, Séverin/Fink, Julia/Dillenbourg, Pierre (2014): "The Dynamics of Anthropomorphism in Robotics." In: Hri 14 Proceedings of 2014 ACM/IEEE International Conference on Human-Robot Interactions, New York: ACM, pp. 226-227.

Levy, David (2008): Love + Sex with Robots. The Evolution of Human-Robot Relationships, New York, London, Toronto, Sydney, New Delhi, Auckland: Harper Perennial.

Levy, David (2012): "The Ethics of Robot Prostitutes." In: Patrick Lin/Keith Abney/George Bekey (eds.), Robot Ethics. The Ethical and Social Implications of Robotics, Cambridge, Massachusetts, London: Oxford University Press, pp. 223-231.

Loh, Janina (2019a): "Responsibility and Robotethics: A Critical Overview." In: Philosophies. Special Issue Philosophy and the Ethics of Technology, 4/58; https://www.mdpi.com/2409-9287/4/4/58/htm.

Loh, Janina (2019b): "What is Feminist Philosophy of Technology? A Critical Overview and a Plea for a Feminist Technoscientific Utopia." In: Janina Loh/Mark Coeckelbergh (eds.*), Feminist Philosophy of Technology.* Stuttgart: J.B. Metzler, pp. 1-24.

Loh, Janina (2020): "Ascribing Rights to Robots as Potential Moral Patients." In: John-Stewart Gordon (ed.), *Smart Technologies and Fundamental Rights. Book series Philosophy and Human Rights*, Brill, pp. 101-126.

Loh, Janina (2022): “Posthumanism and Ethics.” In: Stefan Herbrechter/Ivan Callus/Manuela Rossini/Marija Grech/Megen de Bruin-Molé/Christopher John Müller (eds.), Palgrave Handbook of Critical Posthumanism, Palgrave Macmillan, Cham, Online First Publication; https://doi.org/10.1007/978-3-030-42681-1_34-2.

Loh, Janina (2023, in print): “Are Dating Apps and Sex Robots Feminist Technologies? A Critical-Posthumanist Alternative.” In: Jordi Vallverdú (ed.), Gender in AI and Robotics. Gender Challenges from an Interdisciplinary Perspective, Springer.

Markowitz, Judith A. (ed.) (2015): Robots that Talk and Listen. Technology and Social Impact, Berlin, Boston, München: De Gruyter.

Marr, Bernard (2017): “The 4 Ds Of Robotization: Dull, Dirty, Dangerous and Dear.” In: Forbes, October 16, 2017; https://www.forbes.com/sites/bernardmarr/2017/10/16/the-4-ds-of-robotization-dull-dirty-dangerous-and-dear/?sh=3df1b5a13e0d.

Marsh, Amy (2010): “Love among the objectum sexuals.” In: Electronic Journal of Human Sexuality 13; http://www.ejhs.org/volume13/ObjSexuals.htm.

Misselhorn, Catrin (2009): “Empathy and Dyspathy with Androids: Philosophical, Fictional and (Neuro-) Psychological Perspectives.” In: Konturen 2, pp. 101-123.

Misselhorn, Catrin (2013): “Robots as Moral Agents?“ In: Frank Rövekamp/Friederike Bosse (eds.), Ethics in Science and Society. German and Japanese Views, München: iudicium, pp. 42-56.

Mlot, Stephanie (2018): “Sex Robot Samantha Upgraded With Moral Code.” Originally In: Geek.com; https://www.nexusnewsfeed.com/article/science-futures/sex-robot-samantha-upgraded-with-moral-code/.

Moor, James H. (2006): “The Nature, Importance, and Difficulty of Machine Ethics.” In: Intelligent Systems IEEE 4, pp. 18-21.

Morgan, Rhian (2015): “Looking for robot love? Here are 5 sexbots you can buy right now.” In: METRO News, September 13; https://metro.co.uk/2017/09/13/looking-for-robot-love-here-are-5-sexbots-you-can-buy-right-now-6891378/.

Mori, Masahiro (1970): “Bukimi no tani.” In: Energy 7, pp. 33-35; translated by Karl F. MacDorman/Takashi Minato (2005): “On the Uncanny Valley.” In: Proceedings of the Humanoids-2005 workshop. Views of the Uncanny Valley, Tsukuba, Japan.

Nussbaum, Martha: Frontiers of justice. Disability, nationality, species membership (= The Tanner lectures on human values), Cambridge MA: Belknap Press 2007.

Richardson, Kathleen/Coeckelbergh, Mark/Wakunuma, Kutoma/Billing, Erik/Ziemke, Tom/Gómez, Pablo/Vanderborght, Bram/Belpaeme, Tony (2018): "Robot Enhanced Therapy for Children with Autism (DREAM). A Social Model of Autism." In: IEEE Technology and Society Magazine 37, pp. 30-39.

Scheutz, Michael (2012): "The Inherent Danger of Undirectional Emotional Bonds between Humans and Social Robots." In: Patrick Lin/Keith Abney/George Bekey (eds.), Robot Ethics. The Ethical and Social Implications of Robotics, Cambridge, Massachusetts, London: Oxford University Press, pp. 205-221.

Seibt, Johanna/Noskov, Marco/Schack Andersen, Soren (eds.) (2016): What Social Robots Can and Should Do. Proceedings of Robophilosophy 2016/TRANSOR 2016, Amsterdam: IOS Press.

Shibata, Takanori/Wada, Kazuyoshi (2011): "Robot Therapy. A New Approach for Mental Healthcare of the Elderly – A Mini-Review." In: Gerontology 57, pp. 378-386.

Stasiénko, Jan (2015): "Bizarre marriages. Weddings as a form of legitimization of intimate relations with non-human agents", pp. 80-93; https://www.researchgate.net/publication/308910979_Bizarre_marriages_Weddings_as_a_form_of_legitimization_of_intimate_relations_with_non-human_agents.

Sullins, John P. (2006): "When Is a Robot a Moral Agent?" In: International Review of Information Ethics 6, pp. 23-30.

Tavani, Herman T. (2018): "Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights." In: Information 9; https://www.mdpi.com/2078-2489/9/4/73.

Terry, Jennifer (2010): "Loving Object." In: Trans-Humanities 2, pp. 33-75.

Turkle, Sherry (2005): "Relational Artifacts/Children/Elders. The Complexities of CyberCompanions." In: Proceedings of the Cognitive Science Society Workshop on Android Science, Cambridge MA, pp. 62-73.

Turkle, Sherry (2007): "Authenticity in the age of digital companions." In: Interaction Studies 8, pp. 501-517.

Turkle, Sherry (2010): "In good company? On the threshold of robotic Companions". In: Yorick Wilks (ed.), Close Engagements with Artificial Companions. Key social, psychological, ethical and design issues, Amsterdam: University of Oxford Press, pp. 3-10.

Turkle, Sherry (2011): Alone Together. Why We Expect More from Technology and Less from Each Other, New York: Basic Books.

Wada, Kazuyoshi/Shibata, Takanori/Musha, Toshimitsu/Kimura, Shin (2008): "Robot Therapy for Elders Affected by Dementia." In: IEEE Engineering in Medicine and Biology Magazine July/August, pp. 53-60.

Wallach, Wendel/Colin, Allen (2009): Moral Machines. Teaching Robots Right from Wrong. Oxford, New York: Oxford University Press.

Whitby, Blay (2012): "Do You Want a Robot Lover? The Ethics of Caring Technologies." In: Patrick Lin/Keith Abney/George Bekey (eds.), Robot Ethics. The Ethical and Social Implications of Robotics, Cambridge, Massachusetts, London: Oxford University Press, pp. 233-247.

Złotowski, Jakub/Proudfoot, Diane/Yogeeswaran, Kumar/Bartneck, Christoph (2015): "Anthropomorphism. Opportunities and Challenges in Human-Robot Interaction." In: International Journal of Social Robotics 7, pp. 347-360.

Part I – Understanding, Defining, Conceptualizing

Virtues, Robots, and Good Lives: Who Cares?

Charles M. Ess

1 Introduction

I begin with a general overview of the primary elements of Virtue Ethics (VE) as a global tradition which in its Western*[1] developments turns centrally on *phronēsis*, both "practical wisdom" and the capacity for reflective judgment. Starting with Antigone, *phronēsis* grounds what become Western* traditions of civil disobedience and conscientious objection. At the same time, there are strong arguments that *phronēsis* is not computationally tractable (i.e., *phronēsis's* processes and capabilities cannot be fully replicated by computational technologies and techniques such as Artificial Intelligence (AI)). This will have important consequences for the uptake of VE as an increasingly central framework for Information and Computing Ethics (ICE). This is in part as ICE *qua* global ethics must avoid what I have called "computer-mediated colonization," i.e., the imposition of Western* norms and values upon the rest of the world. A *pluralistic* VE – as I have developed in terms of a *pro hen* interpretive pluralism – seeks to avoid such colonization as well as the fragmentation and potential violence of a simple ethical relativism: that is, as such relativism abandons efforts to achieve shared norms, practices, etc., cultural differences can be used to justify violence in the name of protecting local identities, norms, practices, etc. against all "Others" who may (appear to) threaten such local identities, etc.[2] How privacy is understood across diverse

[1] Following a convention suggested by Grodzinsky, Miller, and Wolf (2008), the asterisk indicates that this term is deeply fraught, contested, and highly ambiguous – perhaps even no longer meaningful at all in certain ways. There is no space here to explore the possible variations and nuances: the term can be used here only as a shorthand or heuristic. I will also flag privacy* in this way for similar reasons.

[2] For example, during their visit to Oslo, Norway, in late January, 2022, when confronted by journalists with questions regarding their plans for protecting young girls' rights to

cultures – first of all as cultures vary their emphases on more *individual* vis-à-vis more *relational* conceptions of selfhood – serves as a primary example of such pluralism.

This second section continues with a series of common objections to VE and replies. Especially important here are critiques of VE as *human-centric*. On the one hand, feminist and posthumanist decenterings of the (male) human are vitally important to enhancing our possible relationships with computational agents as well as the larger world: such decentering overrides especially Christian and Cartesian dualisms entailing master-slave relationships between mind vs. body, male vs. female, human vs. nature as well as machines as artifacts – foregrounding rather an inextricable web of relationality and greater equality with such Others. On the other hand, throwing out modern notions of autonomy entirely thereby ejects the grounding for the still urgent insistence on equality, emancipation, and democratic norms.[3] These tensions thus force the question: How to preserve and enhance democratic rights and norms of equality, emancipation, and disobedience in a posthumanist philosophical anthropology?

As sexuality and intimacy are especially sensitive dimensions of our possible relationships with social robots as driven by Artificial Intelligence (AI) and Machine Learning (ML), in the third section I explore this tension by reviewing an application of VE to sexbots. The feminist phenomenological and ethical analyses of Sara Ruddick (1975) helpfully conjoin virtue ethics with deontological norms of equality and respect as part of "complete sex" – and further uncover both care and loving itself as virtues, i.e., capabilities that must be cultivated and fostered. I argue that while "good sex" between human and

education and women's rights in general, the Taliban spokesman replied that these were Western perspectives that did not apply to them. This use of relativism to defend local differences has been followed up subsequently with sometimes violent repression of these rights in Afghanistan in the name of protecting the Taliban's interpretations of Sharia as the law of the land. Cf. Wong (2020).

[3] As Janina Loh makes clear, not all postmodernists or posthumanists threaten these notions of autonomy (2022). But certainly early postmodernists such as Lyotard ([1979] 1984) undermine these notions as merely parts of the Enlightenment «master narrative», i.e., in contrast with stronger ontological and/or philosophical anthropological claims grounded in philosophical arguments, etc. Reduced to narrative, such claims can then be easily switched out with postmodernist narratives that can elide or simply eliminate these notions. At least some versions of early postmodernism, especially within the US context, thereby fell into simple epistemological, ontological and ethical relativisms that eliminated these central claims entirely.

machines may well be possible and desirable, a "complete sex" is not. As a start, Ruddick's conditions for complete sex include first-person phenomenal consciousness, mutuality of desire, and autonomy as issuing in a deontological insistence on equality and respect. As with *phronēsis*, however, these and related conditions such as genuine emotions and embodied desire, are (likely) not computationally tractable.

At the same time, such differences between human and machine – starting with the absence of autonomy – might inspire us to see these devices as simple means to our ends, treating them analogously to "just meat," as articulated in prominent criticisms of sexbots as, for example, reinscribing traditional patriarchal and racist attitudes. To counter this, I turn in the fourth section to more recent developments in a "more than human" ethics of care. These developments extend a postmodern rejection of the Cartesian master-slave relationship between mind and body, human and nature, male and female, and thus humans and machines. This expanded ethics of care thus add to central notions of a pluralistic VE, as exercised by relational autonomies, to a posthuman philosophical anthropology that promises to foster possibilities of good sex between humans and machines. This is particularly prevalent as machines are now understood to be endowed with a default goodness and moral status. Such an ethics will also help in cultivating still larger relationships of care and repairment between humans and the more than human webs of relationships within which we are inextricably entangled.

In sum: I argue that good lives of flourishing are possible by way of a relational VE coupled with ethical pluralism vis-à-vis four problems, beginning with developing an Intercultural Information Ethics that avoids both ethical relativism and computer-mediated colonization. Second, risks to undermining autonomy as core to democratic polity, norms, and emancipatory imperatives are initially resolved via feminist notions of *relational autonomy*. Third, risks of reinstating master-slave relationships in conjunction with sexbots can be overcome for the sake of "good sex" with our machineries by way of an ethics of care for our more than human webs of relationships. A fourth problem, that of ethical deskilling, of unlearning the virtue of care, can be resolved via a heightened commitment to care. Our cultivation of care, along with *phronēsis*, loving itself, and *courage* thus emerge as at least necessary conditions for good lives of flourishing and pursuits of emancipation, equality, and respect in our human and more than human webs of relationships.

To see how this is so, we begin with a careful look at virtue ethics.

2 Virtue ethics in Information and Communication Technology Ethics

2.1 VE – an overview

Versions of VE are found more or less everywhere – both in major global traditions such as Confucian thought as well as classical Aristotelian VE, but also in the Abrahamic religions (Judaism, Christianity, Islam) as well as indigenous traditions. This global reach may arise in part from the fact that VE starts with a very simple – and, it would seem, a near *universal* human interest: namely, in achieving a life of contentment or *eudaimonia*. This is to say that VE starts with the question: what sort of person do I want/need to become to feel content (*eudaimonia*) and to *flourish*, to feel that my abilities and capacities unfold and become better over time – not simply in the immediate present, but across the course of my entire (I hope, long) life? This basic interest in *a good life* leads to the central insight that certain basic *habits, capacities,* and *abilities*[4] are necessary to acquire and cultivate in order to achieve such contentment and flourishing. To begin with, as Shannon Vallor has foregrounded, the capacities of patience, perseverance, empathy, trust and, indeed, loving itself are necessary conditions for communication, deep friendships, as well as long-term commitments such as partnership or marriage, much less parenting (e.g. Vallor 2010). As I paraphrase her accounts – recall when you were very young and small and were dragged to visit some boring relative or another: how much patience, perseverance, and so on did you have when confronted with the expected communication and relationship with, for example, an ancient grandmother? Such qualities are notoriously absent in four-year-olds,

[4] In several languages and cultures, "virtue" and thereby "virtue ethics" can carry overtones of a sort of moralizing that is no longer useful or attractive. For example, in Anglophone usage "virtue" became a euphemism for a young woman's virginity: to "lose one's virtue," especially in more Victorian and Puritanical times, was a great sin, shame, and scandal. In Scandinavia, my students tell me, when I talk about virtue ethics I sound like a "Konfi-leder," the person leading young people preparing for Confirmation – something like a Sunday School teacher in the US. If "virtue" in your language / cultural experiences carries such off-putting, moralizing overtones, you can substitute "capability" or "ability" for the term. In fact, there is a related – but also importantly different – form of ethics developed by Martha Nussbaum and Amartya Sen titled "capability ethics" (Robeyns/Byskov 2021). But it would go too far afield to explore the important similarities and differences in this context.

for example: as should be clear, these are capacities or abilities that we acquire and develop only over a very long time – indeed, a lifetime. But as, for example, musicians, gamers, people skilled in various crafts, etc. will recognize, we must also acquire and practice these abilities – for instance, when confronting a difficult game, musical score, project, and so on – along with our more particular skills and abilities, in order to accomplish our aims with excellence and thus enjoy *eudaimonia*.

More generally, as Vallor elaborates (2016: pp. 36-49), diverse VE traditions have emerged emphasizing contentment and flourishing in terms of *harmony* [*He* in Confucian thought] – either a more internal harmony (e.g., among the elements of the *psyche* ["soul"] in Plato) or more external harmony with the larger society and (super)natural order (e.g., *tian* ["heaven"] in Confucian thought). Specific to Western* traditions, however, is the central virtue of *phronēsis* – a term that means both "practical wisdom" as well as the essential capacity of a reflective form of *judgment*.

Phronēsis is arguably centrally distinctive to Western* traditions – starting with Sophocles' play *Antigone*. Antigone's brother Polynices has been killed in the Theban civil war and left on the battlefield on orders of Creon, the new tyrant. Antigone's familial and religious obligation is to give her brother a proper burial: doing so, by Creon's decree, will be punished by death. Antigone's own use of *phronēsis* leads to her making the startling – indeed, revolutionary – decision to defy Creon's orders and bury her brother, no matter the cost. This "Antigone moment" (Rockwell F. Clancy, unpublished manuscript: cf. Clancy 2021) appears to root what becomes a distinctively Western* tradition of civil disobedience and protest – as further manifest in Socrates' accusations against Athens, and his willingness to put his obedience to a higher law in conflict with the decrees of the state, even if the cost is death.

In Plato, *phronēsis* is especially tied with the key image and example of the steersman (*cybernetes*): 'An outstanding pilot [*cybernetes*] or doctor is aware of the difference between what is impossible in his art and what is possible, and he attempts the one, and lets the other go; and if, after all, he should still trip up in any way, he is competent to set himself aright' (*Republic* 360e-361a, Bloom trans.; cf. *Republic* I, 332e-c; VI, 489c, in Ess 2007a: 15). *Phronēsis* appears here as both a practical wisdom in the understanding of what is both possible and not possible, and as a capacity for judging what the best course or treatment might be within a specific context. Most importantly, *phronēsis* entails our capacity to learn from our mistakes: experience, including making what turn out to be *judgments* leading to undesirable consequences or errors, can thus inspire an *ethical* self-correction – and thereby enhance and refine both the substance of our practical wisdom as well as our capacity for reflective judgment.

Last but not least, *phronēsis* is especially critical vis-à-vis social robots as driven by AI / ML. As Joseph Weizenbaum's title suggests, *Computer Power and Human Reason: From Judgment to Calculation* (1976), *phronēsis* is arguably distinct from and not tractable by computational techniques. These arguments have been amplified in recent years – perhaps most powerfully by Katharina Zweig, a bio-informaticist who explores in considerable but highly accessible detail what she characterizes as the "machine room" of AI / ML (2019). The first part of her German title is telling: *Ein Algorithmus hat kein Taktgefühl* – an algorithm has no sense of tact, where tact means primarily a sense of what is socially appropriate. Zweig links this sense of tact explicitly with *Urteilskraft* – judgment. She forcefully argues that given the details of how algorithms and ML systems are designed and implemented, they are unable to replicate the sorts of qualitative, context-sensitive judgments that human beings must thereby make in their stead (cf. Cantwell Smith 2019; Sullins, 2021).

2.2 VE in information and computing ethics

To begin with, the *cybernetes*' capacity for self-correction is manifestly the inspiration for 'cybernetics' as taken up by Norbert Wiener as 'the science of messages' ([1950]1954: 77; cf. Bynum, 2010). Wiener's particular vision draws from the French Enlightenment notion of *liberté* – understood more precisely as "...the liberty of each human being to develop in his freedom the full measure of the human possibilities embodied in him" (1954: 106; Ess 2019: 82). Such unfolding of one's best possibilities is manifestly part of a larger pursuit of flourishing and a good life.

The past twenty years or so have witnessed an ever increasing emphasis on VE, for instance, in Kari Gwen Coleman's "Android Arête: Toward a virtue ethic for computational agents" (2001). Coleman hereby begins a now prominent thread of VE especially in conjunction with Artificial Intelligence (AI), Machine Learning (ML), and social robots as incorporating these technologies.

In particular, VE and its central concept of *phronēsis* entail that it is not just professional philosophers who might be good at making reflective judgments: rather, as is emphasized especially in its Aristotelian antecedents, *all* *of us* are "ethicists" precisely in the sense that each of us must make such difficult reflective judgments throughout our lives, in the face of both large and small ethical dilemmas. This insight has become powerfully instantiated in the recent work of the International Electronic and Electrical Engineering (IEEE) association' project to develop "ethically-aligned design" for AI and ML. The executive summary of the first major document in this direction is worth quoting in its entirety:

> "Whether our ethical practices are Western (e.g., Aristotelian, Kantian), Eastern (e.g., Shinto, 墨家/School of Mo, Confucian), African (e.g., Ubuntu), or from another tradition… our goal should be ***eudaimonia***, a practice elucidated by Aristotle that defines human well-being, both at the individual and collective level, as the highest virtue for a society.
>
> …
>
> Autonomous and intelligent systems should prioritize and have as their goal the explicit **honoring of our inalienable fundamental rights and dignity** as well as **the increase of human *flourishing*** and **environmental sustainability**." (Ess 2019: 2, emphases added)

This last sentence shows clear commitments to both deontology as grounding rights and dignity and to the now recognizably central elements of eudaimonia and flourishing. This statement is further helpful as it points out the literally *global* reach of VE traditions, as noted above. This is to say: I (and others) have argued that VE is central, perhaps inevitable in an information and computing ethics (ICE) that is an unavoidably global enterprise. As Krystina Gorniak-Kocikowska (1996) pointed out early on, "computer ethics" must become a global ethics as ICTs become globally distributed technologies, most especially via the internet. Gorniak observed that such a global ICE must meet two criteria. One, such an ethics hence requires recognition as ethically legitimate and compelling in diverse cultures around the world ("global reach"). Two, this ethics must squarely confront and resolve what I have called the problem of "computer-mediated colonization," i.e., how to establish such a global ethics of (quasi-) universal norms, values, principles etc. *without* imposing (colonizing) these *homogenously* across the globe? (Ess 2002a, 2002b, 2006a, 2006b, 2020; Bynum/Kantar 2021)

These issues have been taken up since at least 1990, in what Rafael Capurro designated as Intercultural Information Ethics (1990; cf. Ma 2021) and what more recently has been designated as an Intercultural Digital Ethics (Aggarwal 2020) and Intercultural Ethics of Technology (Verbeek 2021). My approach has focused on the development of what is most elaborately designated as an interpretative *pros hen* ("towards one") pluralism that seeks to hold together both (quasi-) universal norms or values (or, alternatively, what Clifford Christians (2019) has identified as "proto-norms") alongside irreducible differences distinguishing local practices, traditions, and so on. The central insight here is that we can recognize, in at least some important and prominent examples, that such local differences are not necessarily the result of an ethical relativism – i.e., the claim that there is no such thing as (quasi-) universal values that are ethically legitimate over (most) human cultures and histories: rather, there are only local practices and norms that are thereby ethically relevant solely within ("relative to") a given time and place. *Contra* such relativism, pluralism rather argues that these differences result from diverse *interpretations*, applications, and/or understandings of *shared* norms and values. As a first example, privacy*[5] is a norm that is shared across many cultures in one way or another – but sometimes understood or applied in dramatically different ways. To begin with, in contemporary Western* societies, privacy* is understood as a *positive* good and primary right – primarily for *individuals* in Anglophone societies as well as for more *relational individuals* in Scandinavian societies. For example, privacy rights are part and parcel of Human Subjects Protections in Anglophone societies, whose research ethics emphasize obligations to protect individual's privacy and thereby confidentiality by way of practices of informed consent, anonymization, and so on. In Norwegian research ethics, the emphasis is on protecting the privacy of both a given individual *and* that of those who are included in the close relationships of one's *privatlivet* [private life] and *intimsfære* [intimate sphere] (NESH 2006; Ess 2019).

In sharp contrast, traditional societies emphasize especially *relational* conceptions of the self – i.e., where selfhood and identity are entirely defined by one's relationships with others. In Confucian tradition, for example, who we are as spouse, friend, parent, child, sibling, aunt or uncle, a specific vocation, and member of larger communities – social, natural, and, in some worldviews, "supernatural" – constitute the *entirety* of our sense of self: in contrast with modern Western*, especially strongly atomistic conceptions of selfhood,

[5] As noted in ftn. 1, the asterisk serves as a reminder that this term / concept is so contested and ambiguous that its meaning(s) must be much more carefully specified: but as with "Western*, " to further specify privacy* with required detail and nuance would require considerably more space than is available here. Proceed with caution.

there is no such "self" left over if these relationships are compromised or destroyed (Ames/Rosemont 1999; Ess 2011: 17.) Similar understandings are documented in Buddhist societies, as well in Southern African understandings of "Ubuntu" (Paterson 2007: 157-158). In these societies (at least prior to their interacting with and thereby being influenced by contemporary Western* societies), privacy* was understood either as something entirely negative, for example, as something shameful or hidden (Lü 2005) or as a relational privacy such as familial privacy in traditional Thailand (Kitiyadisai 2005). In a pluralistic interpretation, these sharp differences are thus nonetheless clear examples of *interpreting* or applying a shared norm – refracted, as it were, through the lenses of very different assumptions of selfhood and identity, to start with (cf. Hongladarom 2017; Ess 2019, 2020; Ma 2021).

Such a pluralism helps us to avoid an ethical relativism that, in effect, gives up on a shared global ethic and defends instead – sometimes with violence – the legitimacy of a solely local tradition as irreducibly different from others. At the same time, this pluralism allows for a shared set of global norms that, precisely as refracted and applied differently, thereby protects and preserves local cultural identities, practices, etc. Stated differently, pluralism thus avoids imposing norms *homogenously* in all times and places – erasing cultural differences in a (computer-mediated) colonization or imperialism.[6]

2.3 Virtue ethics – objections, replies, clarifications

2.3.1 Individual cultivation?

Of course, VE is not without its limitations and criticisms. A first set of critiques clusters about the assumption that VE rests and focuses solely upon *individual cultivation*. This is to say, as Luciano Floridi argued early on, VE thereby does not scale up – it is not an ethics that can be adopted on a society-wide scale (Floridi 2013: 164-168; cf. Floridi 1999: 41, in Bay 2021: 361).

[6] *Pros hen* interpretive pluralism is manifestly Western* in its origins: this opens up debate as to whether or not, despite its best intentions, it nonetheless risks a conceptual colonization across diverse cultures. I and others have argued that forms of pluralism can be found in both religious and philosophical traditions "East and West" (e.g., Moon 2003; Ess 2020; cf. Hashas 2021): at the same time, others have argued that pros hen pluralism either needs refinement to fully avoid such colonization (Hongladarom 2021) or is better understood as one approach to pluralism that resonates but is not fully identical with, e.g., the Indian dialogical process of Samvād (Gautam/Singh 2021; Ess 2021a).

Floridi's critique, however, is fundamentally countered first of all by the emphasis in contemporary ICE on *relational* forms of VE. This is certainly the case with Shannon Vallor's foundational work on "the techno-moral virtues" as cultivated among human beings as relational (2016): this means, as Bastiaan Vanaker put it more recently, that virtues are in fact learned among a community of peers (2021: 348-349). This is especially so for a VE drawn from Confucian traditions, as both Vallor (2016) and Pak-Hang Wong (2012, 2020) have developed (cf. Bay 2021: 359-360). I have also pointed out that the ancient roots of Western* VE in Antigone, Socrates and Aristotle likewise presume a more relational self: beyond Vallor's work, such a self is articulated in contemporary feminist accounts of *relational autonomy*. As but one example, Andrea Westlund notes that "Some social influences will not compromise, but instead enhance and improve the capacities we need for autonomous agency" (Westlund 2009: 27; in Ess 2019: 78). These understandings that we can be *more free* through our relationships with others – as others, e.g., can introduce us to and encourage us in our cultivation of specific skills and abilities – are especially well exemplified in specific Scandinavian practices and law. For example, the *allemannsretten*, "every one's right" to access to otherwise private property for the sake of camping, picnicking, and so on, is a form of *inclusive property* or Commons that acknowledges shared rights of access coupled with shared responsibilities to *practice* care in accessing these Common spaces (Ess 2019: 79-80). More specifically, David Gunkel (2018) and Mark Coeckelberg (2020) have applied relational versions of VE to the ethics of AI, ML, and social robots.

2.3.2 Naturalistic objections

These begin with *empirical* observations of how people appear to make ethical judgments and decisions, emphasizing that they don't do so by way of VE, and therefore VE is not a viable ethic. So, for example, in her ethnography of Danish robotics engineers, Jessica Sorensson (2019) documents in fine detail how these engineers largely do not think of "ethics" as part of their work – or, more harshly, as simply something that gets in the way. More elaborately, Vanaker (2021: 348-349) summarizes critiques of VE from "situationists" who interpret experiments in moral psychology as demonstrating that people's decisions, for example, to help others, are based more on random environmental factors than on a VE model of moral cultivation. Perhaps most harshly, Vanaker points to recent work in personality studies demonstrating that "both personality and situations guide behavior [such that] what circumstances [and] what traits will affect behavior is the main concern" (2021: 348).

Beyond Vanaker's persuasive rejoinders to these critiques, I would further point out that all of these critiques fit the model of the classic *naturalistic fallacy* – that means, seeking to argue from *what is* the case to what *ought* to be case. As crude examples: most societies have practiced some form of slavery (what is) – therefore, we also ought to practice slavery. Or: most societies have subordinated women and children – therefore we should, too. And so on. In these instances: where our empirical studies show that at least many people do not judge or behave according to VE (what is), therefore, we must argue that people *ought not* to pursue VE. While there is much debate about naturalism in these directions (e.g., Baldwin 2010), I hope the examples of slavery and patriarchy make clear that naturalism is a risky approach to ethics.[7]

2.3.3 VE as (excessively) human-centric?

These criticisms are broadly based within postmodernism and subsequent expressions of feminism, posthumanism, postcolonialism, and, more recently, decolonial analyses and theoretical frameworks. At the risk of a painting with a very broad brush, all raise central critiques of especially "the modern liberal subject" as articulated especially in more atomistic conceptions of a rational self in early modernity. Early postmodernism, for example, undermines these classical assumptions as part of a "master narrative" defining Enlightenment views and aims: *contra* arguments for a reason or rationality that will provide us with a universal knowledge of both *what is* (natural science) and *what*

[7] These examples evoke a further question, however: by contrast, how do we know if / that we are being virtuous when we do what we ought to do? Very briefly: our benchmarks begin with the phronemoi – the good persons whom we recognize as exemplars (equivalent to the *junzi* in Confucian thought) and who serve as sources of guidance as we attempt to (phronetically) judge in a given context / case, e.g., "What would Jesus – and/or Antigone and/or Socrates ... – do?" Other exemplars would include 19th ct. Abolitionists in their struggle to eliminate slavery, the Suffragists to establish women's voting rights, and so on. As these latter examples further suggest, conjoining VE with deontology as insisting on protecting and enhancing our autonomy (as a start) means that those virtues and phronemoi that aim towards greater equality, respect, and emancipation (including, as we will see below, among the more than human webs of relationships) are most likely reliable virtues and moral exemplars indeed (Ess 2016a).

ought to be (ethics and politics) – postmodernism characterizes these as nothing other than relatively arbitrary narratives that can and should be replaced by more contemporary ones (e.g., Lyotard 1979). This inaugurates a broadly shared strategy of "decentering," of moving away from notions of a Cartesian pure reason as defining the human as thereby the "master and possessor of nature" (Descartes [1637] 1972: 119-120; Ess 2017). Similarly, important feminist critiques of early Habermasian notions of the ideal speech situation and the importance of "the unforced force of the better argument" persuasively demonstrated that these notions were excessively "masculinist" and thereby excluded especially women and children as traditionally understood (e.g., Benhabib 1986). Importantly, this strand of critique led not to a final rejection of Habermas, but rather to its revisions and expansions so as to include, e.g., affective dimensions such as empathy and solidarity. More recently, May Thorseth (2008) draws on Iris Marion Young's critique of Habermas' early emphases on a neutral and dispassionate rationality as these exclude the less powerful (traditionally women and children) whose communication may depend upon emotion and rhetoric as well. Coupled with William Regh's defense of rhetoric in a Habermasian model and Kant's account of *judgment* as interwoven with aesthetic and affective dimensions, Thorseth develops an expanded notion of Habermasian notion public deliberation as incorporating *narratives* as well, so as to thereby incorporate especially the voices and experiences of women and children as traditionally portrayed.

Additional critiques along these lines include Donna Haraway's foundational "Cyborg Manifesto" ([1985 1991) as well as more recent postcolonial and decolonial theories (e.g., Battell and Mayblin 2011). In my view, these critiques represent important *extensions* of Enlightenment – indeed, far more ancient – impulses and aims towards emancipation and equality (see especially Ingram 2018). This is manifest in efforts to resolutely emphasize the voices and experiences of those previously marginalized by excessively masculinist notions of reason – women, children, people of color, and so on. Equally importantly, more recent expressions of such decentering seek to further include the non-human or "more than human" within our ethical focus and networks of care. A primary example here is Luciano Floridi's Philosophy of Information (PI) and affiliated Information Ethics. Floridi "… takes reality qua information as intrinsically valuable" – a philosophical position manifestly inspired by the rise of ICTs and networked computing technologies as now "enveloping" us in what Floridi has aptly called "the infosphere" (2014). Floridi's information ethics thus deeply resonates with and takes on board in varying ways any number of shifts over the past several decades within sociology and then environmental and feminist philosophies that share a central focus on, e.g., "webs of relationships" (Gilligan 1982) and the interconnection of all things more broadly (Ess 2009: 161-162). These shifts have been further accelerated by feminist and related critiques of 19th ct. positivism as an extension of Descartes' notion of a masculinist reason radically divorced from body and thereby nature more generally (such a divorce is the philosophical precondition for then claiming that men – meaning *men* – can then assert themselves as masters and possessors of nature: Descartes, *ibid*; Ess 2017). Perhaps most importantly, the rise of relativity theories and then quantum mechanics dramatically undermines both the ontological and epistemological assumptions required for positivism – leading instead to centrally *relational* notions yet once again, such as "the entanglement of matter and meaning" (Barad 2007; cf. Wendt 2015). Not accidently, these notions further reflect the growing influence over these past decades of other non-dualistic philosophies such as Buddhism and Confucian traditions. Floridi's PI in particular, is further allied with a "philosophical naturalism" to be found in Western figures such as Plato and Spinoza (Ess 2009: 94).

On the one hand, especially given our interest here in social robots and related technologies, these moves towards a non-dualistic decentering of a once privileged male rationality, are crucial. They not only have overcome or, more precisely, transformed various forms of high modern human-centric modes of thinking and *feeling* as we pursue ethically central values and goals of greater equality and emancipation (as foundational to democratic polity and norms as well), and ecological inclusiveness: they thereby ground an emphasis on the moral status of the larger order of "things" – including ICTs and devices such as robots. This is to say: Floridi's PI in particular, along with more relational accounts of VE such as articulated by Gunkel (2018) and Coeckelbergh (2020) attribute a 'default goodness' (my term) that presumes some degree of intrinsic value in all existent entities as inextricably interwoven with one another.

On the other hand, several important criticisms of these developments are likewise critical. Most briefly: for all their demonstrated emancipatory, egalitarian and democratic aims and impulses – they run the risk of throwing out the human / ethical agent with the modernist bathwaters, and thereby pulling the ethical and argumentative rugs out from under the feet of those of us who endorse precisely these central Enlightenment impulses and aims.

Within the domain of ICE, Rafael Capurro (2008) criticized Floridi's PI in these directions. Capurro argued that PI's non-androcentric insistence on the intrinsic (if easily overridable) value of all entities will inevitably lead to a potentially disastrous undermining of our emphases on the value and responsibility of human beings as key moral agents (Ess 2008: 93).

More broadly, any number of feminists early on recognized these risks as well – i.e., that eliminating all trace of modernist conceptions of human beings as (relationally) autonomous moral agents thereby likewise threatens feminist insistence on gender equality and thereby full enjoyment of the multiple rights defining democratic polity – however much, of course, these had only until recently been recognized as rights for white heterosexual males. For example, Linda McClain (1999) has argued for a reconstruction of earlier conceptions of privacy that would recognize feminist critiques of the modern liberal subject whilst avoiding its complete deconstruction. The preservation of privacy is especially crucial in the US context as the central legal defense for abortion rights. As we have also argued more recently, such privacy remains especially critical to protecting and fostering LGBTQIA2S +[8] sexualities, identities, practices, and so on (Lagerkvist et al, forthcoming).

In particular, the emergence of feminist notions of *relational autonomy* is inspired in part precisely by the concern that the complete loss of some version of autonomy would thus undermine central arguments for equality, dignity, respect, justice, emancipation, and democracy (e.g., Westlund 2009; Veltman 2014). Specifically, Andrea Veltman and Mark Piper explicitly link Kant's notion of autonomy to VE: "autonomy is one primary good among others that a person needs to live a good life or to achieve human flourishing" (2014: 2).

More broadly, Mireille Hildebrandt has pointed out that autonomy in modern law encodes the central rights to resist, disobey, and contest what others, including police and other governmental authorities may accuse us of, in a court of law (2015: 10). Hildebrandt thus articulates the modern legal defense of what may also include rights to *conscientious objection*. As we have seen, Western* traditions of conscientious objection can be traced to Antigone and then figures such as Socrates. In the modern era, conscientious objection and practices of non-violent civil disobedience have been central to all manner of emancipatory movements, such as the 19th ct. Abolitionists and

[8] Lesbian, Gay, Bisexual, Transgender, Queer and/or Questioning, Intersex, Asexual, Two-Spirit, and plus, in effect, what is yet to be discerned and claimed in for sexualities and identities (my paraphrase).

Suffragettes, and then in the 20th century, for example, Gandhi's campaign for Indian independence, Martin Luther King, Jr.'s struggle to achieve Civil Rights for US people of color, global opposition to the Vietnam conflict, followed by global efforts to achieve equal rights and recognition for the broad spectrum of identities and sexualities currently indicated with LGBTQIA2S+. Preserving autonomy in some form is hence critical as it grounds primary notions of equality and rights, democratic polity and norms, and the central insistence on equality and emancipations that marks the majority of these diverse developments over the past decades.

In the fourth section I will return to the central question forced by these tensions – namely, how to preserve and enhance democratic rights and norms of equality, emancipation, and disobedience in a posthumanist philosophical anthropology? But first, I take up a central application of VE to sexbots as the site of centrally human experiences of sexuality and intimacy. This exploration opens up the possibility of "good sex" with machines – but also ethical risks of treating them as simply means to our own ends. The fourth section will explore a care ethics oriented towards the "more than human" web of relationships as including machines and thereby counters the risks of our treating them as mere devices and slaves.

3 Virtue ethics, sex, and robots

As in many other domains, the ethical proof is in the ethical pudding. That is, what do we gain or achieve by taking up VE with regard to robots – in this case, sexbots? I have taken up virtue ethics, coupled with deontology and intimations of what becomes an ethics of care (Ruddick 1975; Ess 2018), as a primary framework with which to approach the many contentious questions surrounding possibilities of love and sex with robots. I review this application here first for the sake of what it may contribute to our assumptions and understandings of love and sex may be – certainly between humans and possibly between humans and machines. Moreover, this analysis thereby foregrounds important ethical limitations to human-machine sex – ones that at the same time will be helpful as we consider in the final section what it might mean to extend VE and care ethics to "more than human" webs of care that encompass not only our technologies, but the larger (super-)natural orders within which we must all live and breathe.

As we have seen, taking up VE in conjunction with robots is by no means a recent turn (Coleman 1999). Earlier work has also applied deontological ethics – for example, to the matter of a central form of *deception*. On the one hand, enthusiasts for robot sex and love – primarily, David Levy (2007) – argue that even if robots cannot genuinely *feel* love, affection, and/or desire for us, this is no objection or hindrance towards our feeling such for them. Specifically, in what is now well documented and exploited in *affective design* – we humans are easily triggered into feeling care and concern for inanimate devices insofar as they can imitate behaviors indexing emotive states. *Contra* Levy's endorsement of such a design, John Sullins has clearly argued that such design amounts to deceptive trick – where deception is morally objectionable: "It is unethical to play on deep-seated human psychological weaknesses put there by evolutionary pressure as this is disrespectful of human agency" (Sullins 2012: 408; in Ess 2016b: 65).

In turn, I have taken up Sara Ruddick's careful phenomenological accounts of "complete sex" vis-à-vis "good sex" (1975). Most briefly, Ruddick discerns a series of necessary conditions for complete sex between two human beings,[9] starting with the requirement for the presence of a first-person phenomenological consciousness – minimally, a consciousness that is both aware of itself and the larger world, including the beloved, and thereby imbued with both moral agency and critical reflection. Such self-consciousness is specifically interwoven with emotions, including those of love and care, and embodied desire: for Ruddick, these must be genuine emotions and desire, not faked or feigned. According to Ruddick, moreover, we experience complete sex when our desire for one another is *mutual*: this means not simply that our desire for one another will be (roughly) equal – more fundamentally, I desire that the beloved desire my desire. Such mutualities are then key to a first ethical consequence and condition: they encourage us to regard one another as free and thus as equals, thereby meeting the first Kantian requirements for autonomy, namely reciprocal respect and thus equality.

Ruddick is perfectly aware that such conditions are rarely met: among other things, it is notoriously difficult in sexual encounters not to let ourselves be overcome by desire to the point that we regard the beloved as mere

[9] While Ruddick seems to assume heterosexuality between two persons – there is nothing that I can find in her account that would restrict it to either heterosexuality and/or sex/intimacy between to just two persons. But I encourage curious and critical readers to work through her account and decide for themselves.

"meat" – in Kantian terms, simply a means to fulfill our ends.[10] Neither does Ruddick argue that "complete sex," as meeting these conditions, is the only form of "good sex." Rather, good sex – for example, between partners who may experience different levels of desire, but who still hold one another as autonomous persons to be respected – is very likely far more common and thereby still ethically in order. At the same time, however, these challenges to full mutuality means that loving in these ways is a *virtue* – a capability that hardly comes naturally, but must rather be cultivated and practiced.

Given these conditions: is complete sex between human and machine possible? Clearly not. To begin with, for all the advances in recent years in the domains of AI / ML, the "settled position" within these domains that these technologies completely lack any possibility of first-person awareness or self-consciousness (Selmer Bringsjord, personal communication). The same holds true for acquiring genuine emotions and most especially a felt sense of embodied desire (Bringsjord et al. 2015: 2; cf. Searle 2014; Nyholm/Frank 2017). To use a common image: such devices would be zombies – specifically, "moral zombies" (Véliz 2021; cf. Rambukkana 2021; Peeters/Haselager 2021.)

Perhaps most importantly: as we saw above, the key virtue of *phronēsis* is not computationally tractable. Such a capacity is required, however, first of all to *interpret* the contexts and determine what, if any, particular gesture, motion, etc. is indeed desired and desirable; what of these work to sustain fundamental respect for the beloved as an equal and thereby sustain the autonomy of the other. While there might be machine approximations (i.e., Sullins' "artificial *phronēsis*," 2021), the fact remains that a datafied *phronēsis*, like artificial desire and faked emotions in machine thereby lacks the fundamental condition of autonomy that undergirds phronetic judgment.

This is not to say, however, that sex with robots – at least with full awareness and acceptance of these limitations – would *prima facie* be ethically problematic: on the contrary, it is easy to see that sex with robots could be good sex in a number of contexts, including therapy, aiding those whose disabilities

[10] To be precise: one formulation of Kant's Categorical Imperative states: "act so that you treat humanity, whether in your own person or in that of another, always as an end and never as a means only" [Kant [1785] 1959, p. 47]. Our lives manifestly require that we treat one another means – such as the store cashier who serves as a means to expedite my purchases. But I am still fully obligated to treat the cashier as an autonomy due full respect and equality – i.e., never as a means only. By the same token, both good and complete sex entail sustaining respect for the Beloved as an autonomy, never as a means or mere object only.

challenge the possibilities of establishing intimate relationships with others, and so on.

Certainly, a raft of additional objections to robot sex are important to notice and take on board, starting with how the sexbot industry caters to and thus inscribes and reinforces highly patriarchal assumptions regarding who (what) is sexually attractive and arousing for Western* cis heterosexual men – namely, young, if not childlike women, Asian women, and so on (Richardson: 2015). These conceptions are further manifest in what Mia Consalvo aptly identifies as the "techno-femme fatale" in science fiction from Fritz Lang's *Metropolis* through any number of more contemporary films and TV series (2004). I have argued elsewhere that these conceptions appear to be rooted in not only Cartesian dualisms that divorce a pure (male) mind from an impure (female) body – and with it, sexuality and nature more broadly. Moreover, such a Cartesian mastery and possession of nature reflects a master-slave conception of the relationship between men and women in turn (Ess 2017).

Most importantly here: *contra* these backgrounds and prevailing designs of sexbots, Ruddick's account of good sex and complete sex would rather endorse our treating sexbots and related machineries more broadly as something more than the technological equivalents of "just meat," as sex slaves that are solely means to our own ends. But especially as machines will lack autonomy and related capabilities such as real emotion, embodied desire, and *phronēsis* – Ruddick's interweaving of care ethics with deontology provides minimal, if any grounds, for countering master-slave relationships between humans and machines, as deontological duties of respect and norms of equality turn on autonomy. Good sex – i.e., as sustaining relationships of respect rather than exploitation – between humans and machines will require an additional ethical layer.[11]

[11] As noted earlier, good or ethically legitimate sex with robots is certainly possible, and there may be specific exceptions to this general ethos of approaching robots with postures of care and respect, such as particular sorts of therapies, providing aid to those whose disabilities challenge their possibilities for developing intimate relationships with humans, and so on. But see also Sparrow (2017) and Danaher (2017) as important contributions to the debates over what may and may not be justified regarding our treatment of robots, e.g., as objects of (stimulated) rape.

4 Preserving human autonomy / democracy + moral status of robots (and beyond): the "more than human" ethics of care

We have seen above a "relational turn" as part of the unfolding of contemporary VE, including recent ethics work on social robots (Gunkel 2018, Coeckelbergh 2020; Ess 2021b; cf. Bynum/Kantar, 2021). This relational turn is in part a response to critiques of VE as both (excessively) individualist as well as intrinsically human-centric (e.g., starting with postmodernism). At the same time, relationality along with some ways of deconstructing or decentering the modern liberal (especially male) subject run the risk of thereby throwing out notions of autonomy that are central to ethical arguments for respect, equality, democratic polity, and thereby emancipation of the countless "others" who have otherwise been marginalized and exploited in patriarchal authoritarian hierarchies. These tensions led us to our second problem: How to preserve and enhance democratic rights and norms of equality, emancipation, and disobedience in a posthumanist philosophical anthropology?

I initially foregrounded feminist notions of relational autonomy as conceptions that are consonant with the many maneuvers of deconstruction, decentering, decolonization, and so on – while nonetheless preserving understandings of autonomy that can sustain equality, democracy, and emancipation. We can now expand on these understandings in a final way by turning to recent work in ethics of care by Christina Mörtberg (2021).

In her keynote address "Work, Place, Mobility and Embodiment: 'Recovery' or Repairment in a Covid and Eventually Post-Covid World?" Mörtberg foregrounds a "more than human" ethics of care that is initially rooted in the care ethics of Joan Tronto (1993). Especially as developed by Maria Puig de la Bellacasa (2012, 2017), care is extended beyond the human-centric approaches first articulated in Gilligan (1982), Sara Ruddick (1980) and Nel Noddings (1984). Puig de la Bellacasa specifically expands notions of human-centric care (and virtue) ethics to encompass the larger domains of robots, technologies, and the larger natural (and, for some, "supernatural") worlds. This issues in a decentered ethics of care that understands "the circulation of care as everyday maintenance of the more than human web of life, conceived as a decentered form of vibrant ethicality, as an ethos rooted in obligations made necessary to specific relations …" one that further emphasizes "care as a doing," not simply a moral intention (Puig de la Bellacasa 2017: 219; in Mörtberg 2021). Mörtberg further connects this decentered ethic with "repairment":

> "it is … a species activity that includes everything that we do to maintain, continue, and repair our "world" so that we can live in it as well as possible. That world includes our bodies, our selves, and our environment, all of which we seek to interweave in a complex, life-sustaining web." (Fisher/Tronto 1991: 40 in Tronto 1993: 103; Mörtberg 2021)

Mörtberg further draws on Haraway's understanding of situated knowledges (1988): reflecting both developments within natural science such as relativity theories and Quantum Mechanics, which emphasize (following Kant, in fact) that knowledge is relative to and dependent upon the observer, Haraway further invokes feminist critiques arguing that knowledge is always partial, local / located, and thereby situated. This is by no means epistemological relativism – but rather issues in a knowledge *pluralism* that argues for a multiplicity of ways of knowing that may be different but complimentary to one another.

Last but not least: Shannon Vallor has extensively explored how care – specifically in the context of carebots – is a *virtue*, that means, a capacity or ability that must be fostered, "…an activity of personally meeting another's need, one that, if properly habituated and refined into a practice, can also become a manifestation of personal excellence" (2016: 221). Care is thus often simple hard work: "We must *learn* [practice] how to care in the right ways, at the right times and places, and for the right people" (*ibid*). The promise of carebots is so compelling in part as we can thereby offload the burdens of care to the machines. But whatever advantages such offloading might offer, to hand over care entirely would mean what Vallor elsewhere discusses as de-skilling, as the loss of critical virtues. That is, such offloading would thereby reduce or eliminate the need for and contexts in which we practice care as a virtue – such that our capabilities for caring improve and become more habitual. In this case, "…people who care for others *well* are among those examples of human excellence that we recognize and respect most readily" (*ibid*). While not all of us will emerge as such exemplars (*phronemoi*) of care, our acquiring and cultivating care is manifestly central to our pursuing good lives of flourishing, as co-relational human beings in (close) relationships with (a few) Others in these "more than human" webs of relationships.

5 Virtues, Robots and Good Lives: Who Cares?

We now come around full circle, returning to the figure of Antigone: we can now see that her *care* and duties of care to her brother Polynices exemplify care as a virtue – one that is further dependent on *phronēsis* for determining within the cauldron of deeply conflicting norms and desires (to obey Creon and save her life, or to care for her brother in ways that *in this context* will result in her death) what choice she finally makes. Antigone is in this way a moral exemplar – a *phronemos* whose specific acts and choices stand as specific examples, helping to illustrate what a good person *judges* is the right thing to do under specific circumstances vis-à-vis multiple possibilities.

Antigone's care is manifestly directed first to her brother: but we've also seen that care – along with human agency and autonomy – by no means must be restricted solely to other humans or perhaps sentient beings more generally. On the contrary, any number of classical and contemporary frameworks help us move beyond a human-centric ethics to a "more than human" ethics of care: one that at the same time, at least via conceptions of relational autonomy, help us take on board several decades' worth of vital criticisms of excessively human-centric (if not simply masculinist and patriarchal) conceptions – but without throwing out human agency and autonomy as the root of still central emphases on equality, respect, and emancipation. Especially where emancipation in the modern and contemporary worlds has required the *courage* of conscientious objection and the modern right to dissent, contest, and disobey (Hildebrandt 2015: 10) – Antigone remains a primal exemplar indeed.[12]

We move beyond the human-centric precisely as we extend care to the more than human – recognizing a default, intrinsic worth or minimal goodness of all about us. This move beyond high modern conceptions is at once, however, a recovery of pre-modern sensibilities and understanding, including relationality and non-dualistic ontologies that counter the modern Cartesian

[12] Clive Hartfield (2021) offers an exceptionally careful analysis of Edward Snowden's actions and decisions vis-à-vis three of Shannon Vallor's technomoral virtues – including «courage, with its related virtues of hope, perseverance, and fortitude» (379), arguing that these decisions and actions manifest these virtues. Hartfield does not use the notion of conscientious objection per se, but I see this analysis as strongly supporting an understanding of Snowden as a conscientious objector in the traditions begun with Antigone.

divorce between mind and matter. Whether we do so via traditional Buddhism, Western* philosophical naturalism, Reform Jewish interpretations of *tikkun olam*, and/or a care ethics oriented towards sustaining and cultivating the complex webs of life surrounding us – the message is that pursuing good lives of flourishing entails positive obligations and actions that will take care of and *repair* the world.[13]

Specifically with regard to social robots, this broad framework endorses minimal attitudes of respect and care – whether or not these devices approach some form of autonomy, much less *phronēsis*, as modernist bases for moral status and equality. Such a framework would meet some of the important critiques of sexbots, e.g., as reinforcing the objectification of women, young girls, even children (Richardson 2015) and help foster instead a posture of care and respect that would ideally avoid rendering the machine purely into a means for the human's own ends – i.e., a slave. In this way – and likely others to be developed as the machines themselves develop – sex with a robot could count as good sex, in Ruddick's term.

While sexbots remain an exotic technology still under development, AI / ML in conjunction with biometric devices are rapidly expanding dimensions of our contemporary socio-technical infrastructure. This turn towards a more than human ethics of care has in fact been taken up in a contemporary research project exploring the ethical and especially existential challenges and their possible resolutions (Lagerkvist et al. forthcoming). While this work is still preliminary, it suggests specific and vital directions for applying such a comprehensive ethics of care alongside more familiar frameworks of VE and deontology.

At least if we care to.

[13] *Tikkun olam*, "repairing the world," is understood in Reform Judaism as an impulse towards, e.g., social and environmental justice. https://en.wikipedia.org/wiki/Tikkun_olam. As we have further explored, Scandinavian Creation Theologies likewise endorse an intrinsic goodness to Creation – which, we and others (Foerst 2020) would extend to robots and AI / ML systems: for more specific consequences of these approaches, see Balle and Ess (forthcoming).

Acknowledgements

I first of all want to thank Wulf Loh, Janina Loh, Regina Amnicht Quinn, and Jessica Heesen for their generous hospitality and most fruitful collegiality during my visits at the IZEW (International Center for Ethics in the Sciences and Humanities) – and most certainly for the invitation to offer the keynote lecture for the EMOBOTS conference (May 13-14, 2021). I am likewise very grateful to the conference presenters and participants for their contributions and lively discussions. I am especially grateful to Mark Coeckelbergh (Vienna) and David Gunkel (Northern Illinois University) for most helpful conversations on relationality and ethics vis-à-vis robots and avatars. Last but certainly not least, a thousand thanks to Christina Mörtberg (Linnæus University, Sweden), most especially for her Keynote address to the Spring Workshop of the IFIP WG9.8 (Gender, Diversity, and IT), April 16, 2021, "Thinking with Care as the Transition from Recovery to Repairment".

References

Aggarwal, Nikita (2020): "Introduction to the Special Issue on Intercultural Digital Ethics." In: Philosophy and Technology 33, pp. 547-550.

Baldwin, Tom (2010): "George Edward Moore." In: Edward N. Zalta (ed.), The Stanford Encyclopedia of Philosophy (Summer 2010 Edition); https://plato.stanford.edu/archives/sum2010/entries/moore.

Balle, Simon/Ess, Charles (Forthcoming): "Robotics, Ethics, and Religion". In: H. Campbell/Pauline Hope Cheong (eds.), The Oxford Handbook of Digital Religion.

Barad, Karen Michelle (2007): "Meeting the universe halfway: quantum physics and the entanglement of matter and meaning." Durham, N.C.: Duke University Press.

Bay, Morten (2021): "Four Challenges to Confucian Virtue Ethics in Technology." In: *The Journal of Information, Communication and Society* 19/3, pp. 358-373.

Benhabib, Seyla (1986): Critique, Norm, and Utopia: A Study of the Foundations of Critical Theory, New York: Columbia University Press.

Bringsjord, S./Licato, J./Govindarajulu, N.S./Ghosh, R./Sen, A. (2015): "Real Robots that Pass Human Tests of Self-Consciousness". In: Proceedings of RO-MAN 2015 (The 24th International Symposium on Robot and Human

Interactive Communication). August 31–September 4, 2015. Kobe, Japan, pp. 498-504.

Bynum, T.W./Kantar, Nesiben (2021): "Global ethics for the digital age — flourishing ethics." In: Journal of Information, Communication, and Society 19/3, pp. 329-344.

Bynum, Terrell Ward (2010): "The Historical Roots of Information and Computer Ethics." In: Luciano Floridi (ed.), The Cambridge Handbook of Information and Computer Ethics, Cambridge: Cambridge University Press, pp. 20-38.

Cantwell Smith, Brian (2019): The Promise of Artificial Intelligence: Reckoning and Judgment, Cambridge MA: MIT Press.

Capurro, Rafael (2008): "On Floridi's metaphysical foundation of information ecology." In: Ethics and Information Technology 10/2-3, pp. 167-173.

Christians, Clifford G. (2019): Media Ethics and Global Justice in the Digital Age, Cambridge: Cambridge University Press.

Christman, John (2004): "Relational Autonomy, Liberal Individualism, and the Social Constitution of Selves." In: Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 117/1-2, pp. 143-164.

Clancy, Rockwell F. (2021): "The Merits of Social Credit Rating in China? An Exercise in Interpretive *Pros Hen* Ethical Pluralism". In: Journal of Contemporary Eastern Asia, 20/1, pp. 102-119.

Coeckelbergh, Mark (2020): "How to Use Virtue Ethics for Thinking About the Moral Standing of Social Robots: A Relational Interpretation in Terms of Practices, Habits, and Performance." In: International Journal of Social Robotics 13, pp. 31-40.

Coleman, Kari Gwen (2001): "Android Arête: Toward a Virtue Ethic for Computational Agents". In: Ethics and Information Technology 3/4, pp. 247-265.

Consalvo, Mia (2004): "Borg Babes, Drones, and the Collective: Reading Gender and the Body in Star Trek". In: Women's Studies in Communication 27/ 2, pp. 177-203.

Danaher, John (2017): "Robotic Rape and Robotic Child Sexual Abuse: Should they be criminalized?" In: Criminal Law and Philosophy 11/1, pp. 71-95.

Descartes, René (1972 [1637]): "Discourse on method." In: The philosophical works of Descartes, trans. E. S. Haldane and G. R. T. Ross, Vol. I, Cambridge: Cambridge University Press, pp. 81-130.

Ess, Charles (1995): "Reading Adam and Eve: Re-Visions of the Myth of Woman's Subordination to Man". In: Marie M. Fortune/Carol J. Adams (eds.), Violence Against Women and Children: A Christian Theological Sourcebook, New York: Continuum Press, pp. 92-120.

Ess, Charles (2002a): "Computer-Mediated Colonization, the Renaissance, and Educational Imperatives for an Intercultural Global Village." In: Ethics and Information Technology IV/1, pp. 11-22. Reprinted in: John Weckert (ed.), Computer Ethics. (The International Library of Essays in Public and Professional Ethics.), Hampshire (UK): Ashgate, 2007.

Ess, Charles (2002b): "Liberation in cyberspace . . . or computer-mediated colonization? / Liberation en cyberspace, ou colonisation assistee par ordinateur?" In: *Electronic Journal of Communication/La Revue Electronique de Communication*, 12/3&4.

Ess, Charles (2004): "Beyond *Contemptus Mundi* and Cartesian Dualism: Western Resurrection of the BodySubject and (re)New(ed) Coherencies with Eastern Approaches to Life/Death." In: Günter Wollfahrt/Hans Georg-Moeller (eds.), Philosophie des Todes: Death Philosophy East and West, Munich: Chora, pp. 15-36.

Ess, Charles (2006a): "Du colonialisme informatique à un usage culturellement informé des TIC." In: J. Aden (ed.), De Babel à la mondialisation: apport des sciences sociales à la didactique des langues, Dijon : CNDP - CRDP de Bourgogne, pp. 47-61.

Ess, Charles (2006b): "From Computer-Mediated Colonization to Culturally-Aware ICT Usage and Design." In: P. Zaphiris/S. Kurniawan (eds.), Advances in Universal Web Design and Evaluation: Research, Trends and Opportunities, Hershey, PA: Idea Publishing, pp. 178-197.

Ess, Charles (2008): "Luciano Floridi's philosophy of information and information ethics: Critical reflections and the state of the art." In: Ethics and Information Technology 10, pp. 89-96.

Ess, Charles (2009): "Floridi's Philosophy of Information and Information Ethics: Current Perspectives, Future Directions," special issue on "The Philosophy of Information, its Nature and Future Developments" edited by Luciano Floridi, In: The Information Society 25, pp. 159-168.

Ess, Charles (2010): "The Embodied Self in a Digital Age: Possibilities, Risks, and Prospects for a Pluralistic (democratic/liberal) Future?" In: Nordicom Information 32/2, pp. 105-118.

Ess, Charles (2016a): "Ethical Approaches for Copying Digital Artifacts: What Would the Exemplary Person (*Junzi*) / a Good Person [*Phronemos*] Say?" In: R. Schmücker/D. Hick (eds.), The Aesthetics and Ethics of Copying, . London: Bloomsbury, pp. 295-313.

Ess, Charles (2016b): "What's love got to do with it? Robots, sexuality, and the arts of being human." In: M. Nørskov (ed.), Social Robots: Boundaries, Potential, Challenges, Farnham, Surrey, England: Ashgate, pp. 57-79.

Ess, Charles (2017): "God Out of the Machine?: The Politics and Economics of Technological Development." In: A. Beavers (ed.), Macmillan Interdisciplinary Handbooks: Philosophy, Farmington Hills, MI: Macmillan Reference, pp. 83-111.

Ess, Charles (2019): "Intercultural privacy: a Nordic perspective." In: Hauke Behrendt/Wulf Loh/Tobias Matzner/Catrin Misselhorn (eds.), Privatsphäre 4.0: Eine Neuverortung des Privaten im Zeitalter der Digitalisierung, Stuttgart: J.B. Metzler, pp. 73-88.

Ess, Charles (2020): "Interpretative *pros hen* pluralism: from computer-mediated colonization to a pluralistic Intercultural Digital Ethics." In: Philosophy and Technology 33, pp. 551-569.

Ess, Charles (2021a): "2021 Summer Issue Special: Introduction to JCEA." In: Journal of Contemporary Eastern Asia 20/1, pp. 89-101.

Ess, Charles (2021b): "Towards an Existential and Emancipatory Ethic of Technology." In: Shannon Vallor (ed.), Oxford Handbook of Philosophy and Technology. (online edn, Oxford Academic, 10 Nov. 2020), pp. 588-608.

Fisher, Berenice/Joan C. Tronto (1990): "Toward a feminist theory of caring." In: Emily Abel/Margaret Nelson (eds.), Circles of care: Work and identity in women's lives, Albany: State University of New York Press, pp. 36-54.

Floridi, Luciano (2013): The Ethics of Information, Oxford: OUP.

Floridi, Luciano (2014): The Fourth Revolution: How the Infosphere is Reshaping Human Reality, Oxford: OUP.

Foerst, Annie (2020): "Loving robots? Let yet another stranger in." In: Love, Technology and Theology, London: Bloomsbury Publishing Plc, p. 16.

Gautam, Ayesha/Singh, Deepa (2021): "Building Bridges: Eurocentric to Intercultural Information Ethics." In: Journal of Contemporary Eastern Asia 20/1, pp. 151-168.

Gilligan, Carol (1982): In a Different Voice: Psychological Theory and Women's Development, Cambridge, MA: Harvard University Press.

Gonzalez-Franco, Mar, Slater, Mel, Birney, Megan E., Swapp, David, Haslam, Alexander, Reicher, Stephen D. (2018): "Participant concerns for the Learner in a Virtual Reality replication of the Milgram obedience study." In: PLoS ONE 13/12, e0209704.

Gorniak-Kocikowska, Krystyna (1996): "The Computer Revolution and the Problem of Global Ethics." In: Science and Engineering Ethics 2, pp, 177-190.

Grodzinsky, Frances S./Miller, Keith/Wolf, Marty J. (2008): "The ethics of designing artificial agents." In: Ethics and Information Technology 10, pp. 115-121.

Gunkel, David (2012): The Machine Question, Cambridge, MA: MIT Press.

Gunkel, David (2018): Robot Rights, Cambridge, MA: MIT Press.

Haraway, Donna (1988): "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." In: Feminist Studies 14/3, pp. 575-599.

Haraway, Donna (1991 [1985]): "A cyborg manifesto: Science, technology, and socialist-feminism in the late twentieth century." In: Simians, cyborgs and women: The reinvention of nature, New York: Routledge, pp. 149-181.

Harfield, Clive (2021): "Was Snowden virtuous?" In: Ethics and Information Technology 23, pp. 373-383.

Hashas, Mohammed (ed.) (2021): Pluralism in Islamic Contexts - Ethics, Politics and Modern Challenges, Cham, Switzerland: Springer.

Hildebrandt, Mireille (2015): Smart Technologies and the End(s) of Law, Cheltenham: Edward Elgar.

IEEE (2019): "Ethically Aligned Design, 1st ed."; https://ethicsinaction.ieee.org/.

Ingram, James D. (2018): "Critical Theory and Postcolonialism." In: Peter Gordon/Espen Hammer/Axel Honneth (eds.), The Routledge Companion to the Frankfurt School, London: Routledge, pp. 500-513.

Kant, Immanuel ([1785] 1959): Foundations of the Metaphysics of Morals, trans. Lewis White Beck. Indianapolis: Bobbs-Merrill.

Kitiyadisai, Krisana (2005): "Privacy rights and protection: foreign values in modern Thai context." In: Ethics and Information Technology 7, pp. 17-26.

Lagerkvist, Amanda/ Tudor, Matilda/ Smolicki, Jacek/Ess, Charles M./ Eriksson, Jenny Lundström/Rogg, Maria (forthcoming): "Body Stakes: An Existential Ethics of Care in Living with Biometrics and AI," In: AI & Society.

Levy, David (2007): Love and Sex with Robots: The Evolution of Human-Robot Relationships, New York: Harper Perennial.

Loh, Janina (2022): "Posthumanism and Ethics" In: Stefan Herbrechter/Ivan Callus/Manuela Rossini/Marija Grech/Megen de Bruin-Molé/Christopher

John Müller (eds.), Palgrave Handbook of Critical Posthumanism, Cham: Palgrave Macmillan.

Lowman, Emma Battell/Mayblin, Lucy (2011): "Theorising the Postcolonial, Decolonising Theory." In: Special issue, Studies in social & political thought 19.

Lü, Yao-Huai (2005): "Privacy and Data Privacy Issues in Contemporary China." In: Ethics and Information Technology 7, pp. 7-15.

Lyotard, Jean-François (1984 [1979]): The Postmodern Condition: A Report on Knowledge, Minneapolis: University of Minnesota Press.

Ma, Yuanye (2021): "Language and Intercultural Information Ethics Concepts: A Preliminary Discussion of Privacy." In: Data and Information Management 5/1, pp. 159-166.

McClain, Linda C. (1999): "Reconstructive Tasks for a Liberal Feminist Conception of Privacy." In: 40 Wm. & Mary Law Review 759 3/4, pp. 759-794.

Moon, J. Donald (2003): "Pluralisms compared." In: R. Madsen/T. B. Strong (eds.), The many and the one: Religious and secular perspectives on ethical pluralism in the modern world, Princeton: Princeton University Press, pp. 343-359.

Mörtberg, Christina (2021): "Thinking-with Care: Transition from Recovery to Repairment.", Keynote lecture, IFIP WG9.8 Workshop, "Work, Place, Mobility and Embodiment: «Recovery» or Repairment in a Covid and Eventually Post-Covid World?, April 16, Linköping, Sweden.

NESH (The Norwegian National Committee for Research Ethics in the Social Sciences and the Humanities) (2006): Forskningsetiske retningslinjer for samfunnsvitenskap, humaniora, juss og teologi [Research ethics guidelines for social sciences, the humanities, law and theology]; https://www.etikkom.no/globalassets/documents/english-publications/guidelines-for-research-ethics-in-the-social-sciences-law-and-the-humanities-2006.pdf.

Noddings, Nel (1984): Caring: A Feminine Approach to Ethics and Moral Education, Berkeley: University of California Press.

Nyholm, Sven/Frank, Lily Eva (2017): "From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible?" In: John Danaher/Neil McArthur (eds.), Robot Sex: Social and Ethical Implications, Cambridge, MA: MIT Press, pp. 219-243.

Paterson, Barbara (2007): "We Cannot Eat Data: The Need for Computer Ethics to Address the Cultural and Ecological Impacts of Computing." In: S.

Hongladarom/C. Ess (eds.), Information Technology Ethics: Cultural Perspectives, Hershey PA: Idea Group Reference, pp. 153-168.

Peeters, Anco/Haselager, Pim (2021): "Designing Virtuous Sex Robots." In: International Journal of Social Robotics 13, pp. 55-66.

Puig de la Bellacasa, María (2012): "'Nothing Comes Without Its World': Thinking with Care." In: The Sociological Review 60/2, pp. 197-216.

Puig de la Bellacasa, María (2017): Matters of care: speculative ethics in more than human worlds, Minneapolis: University of Minnesota Press.

Rambukkana, Nathan (2021): "Robosexuality and its Discontents." In: N. Rambukkana (ed.), Intersectional Automations: Robotics, AI, Algorithms, and Equity, Lanham, MD: Lexington Books.

Richardson, Kathleen (2015): "The Asymmetrical 'Relationship': Parallels Between Prostitution and the Development of Sex Robots." In: SIGCAS Computers & Society 45/3, pp. 290-293.

Robeyns, Ingrid/Fibieger Byskov, Morten (2021): "The Capability Approach", In: Edward N. Zalta (ed.), The Stanford Encyclopedia of Philosophy; https://plato.stanford.edu/archives/win2021/entries/capability-approach/.

Ruddick, Sara (1975): "Better Sex." In: Robert Baker/Frederick Elliston (eds.), Philosophy and Sex, Amherst, NY: Prometheus Books, pp. 280-299.

Ruddick, Sara (1980): "Maternal Thinking." In: Feminist Studies 6/2, pp. 342-367.

Sea rle, John (2014): "What Your Computer Can't Know." In: New York Review of Books; https://www.nybooks.com/articles/2014/10/09/what-your-computer-cant-know/.

Sorenson, Jessica (2019): "Toward a pragmatic and social engineering ethics: Ethnography as provocation." In: Paladyn Journal of Behavioral Robotics 10, pp. 207-218.

Sparrow, Robert (2017): "Robots, Rape, and Representation." In: International Journal of Social Robotics 9, pp. 465-477.

Sullins, John P. (2021): "Artificial Phronesis: What It Is and What It Is Not." In: Emanuele Ratti/Thomas A. Stapleford (eds.), Science, Technology, and Virtues: Contemporary Perspectives.

Tronto, Joan C. (1993): Moral boundaries: a political argument for an ethic of care, London: Routledge.

Vallor, Shannon (2010): "Social networking technology and the virtues." In: Ethics and Information Technology 12, pp. 157-170.

Vallor, Shannon (2016): Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting, Cambridge, MA: MIT Press.

Vanaker, Bastiann (2021): "Virtue Ethics, Situationism and Casuistry: Toward a Digital Ethics Beyond Exemplars." In: Journal of Information, Communication and Ethics in Society 19/3, pp. 345-357.

Véliz, Carissa (2021): "Moral zombies: why algorithms are not moral agents." In: AI & Society 36, pp. 487-497.

Veltman, Andrea (2014): "Autonomy and Oppression at Work." In: Andrea Veltman/Mark Piper (eds.), Autonomy, Oppression and Gender, Oxford: OUP, pp. 280-300.

Veltman, Andrea/Piper, Mark (2014): "Introduction." In: Andrea Veltman/Mark Piper (eds.), Autonomy, Oppression and Gender, Oxford: OUP, pp. 1-11.

Westlund, Andrea (2009): "Rethinking Relational Autonomy." In: *Hypatia* 24/4, pp. 26-49.

Wiener, Norbert (1950): The Human Use of Human Beings: Cybernetics and Society, New York: Houghton Mifflin.

Wong, Pak-Hang (2012): "Dao, harmony and personhood: Towards a Confucian ethics of technology." In: *Philosophy & Technology* 25/1, pp. 67-86.

Wong, Pak-Hang (2020): "Cultural Differences as Excuses? Human Rights and Cultural Values in Global Ethics and Governance of AI." In: Philosophy and Technology 33, pp. 705-715.

Zweig, Katharina (2019): Ein Algorithmus hat kein Taktgefühl. Wo Künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können, München: Heyne.

The Relational Turn
Thinking Robots Otherwise

David J. Gunkel

1 Introduction

Ethics is an exclusive undertaking. In confronting and dealing with others—whether another human person, a non-human animal, or an artifact—we inevitably make a decision between *who* is worthy of consideration and respect and *what* remains a mere thing that can be used as we see fit. These decisions are often accomplished and justified on the basis of some fundamental and intrinsic property that is determined to belong to the entity by its very nature. "The standard approach to the justification of moral status is," as Mark Coeckelbergh (2012: 13) explains, "to refer to one or more (intrinsic) properties of the entity in question, such as consciousness or the ability to suffer. If the entity has this property, this then warrants giving the entity a certain moral status."

This way of proceeding has been successfully utilized on both sides of the debate concerning the moral status of AI, robots, and other artifacts. On the one side, those opposing any form of moral status for artifacts assert that these technologies are just things or objects that do not possess and will not come to possess the necessary conditions or capabilities to be considered something more. On the other side, there are those who favor extending some aspect of moral status to AI and robots by arguing that these technological things either have or will soon be able to possess one or more of the necessary and essential properties to be something other than a mere thing. What is interesting about this debate is not what makes the one side different from the other; what is interesting is what both sides already agree upon and share in order to come into conflict in the first place. And the real problem is not that this shared moral scaffolding has somehow failed to work in the face or the faceplate of AI and robots. The problem is that it has and continues to work all-too-well, exerting its influence and operations almost invisibly and without question.

This chapter is designed to respond to this problem. It begins by first identifying and critically examining three seemingly intractable philosophical difficulties with the standard method for deciding questions of moral status. In response to these demonstrated difficulties, the second section will introduce and describe an alternative model, one which shifts the emphasis from internal properties of the individual entity to extrinsic social circumstances and relationships. The final section will then consider three possible objections to this "relational turn" and provide responses to these criticisms. The goal in all of this is not to complicate things but to introduce and formulate a metaethical theory that is more agile in its response to the unique opportunities and challenges of the 21st century.

2 The Properties Approach

In responding to others (and doing so responsibly), we typically need to distinguish between *what* is a thing and *who* is another person. As Immanuel Kant (2012: 40) once described it: "Beings whose existence rests not indeed on our will but on nature, if they are non-rational beings, still have only a relative worth, as means, and are therefore called *things*, whereas rational beings are called *persons*, because their nature already marks them out as ends in themselves, i.e., as something that may not be used merely as means, and hence to that extent limits all choice (and is an object of respect)." This just sounds intuitively correct. We go out into the world and deal with others, knowing there's a difference between other persons who are subject to respect as ends in themselves and those things that are mere objects with instrumental value as a means to an end. As Robert Esposito (2015: 1), who arguably wrote the book on this matter, explains: "If there is one assumption that seems to have organized human experience from its very beginnings it is that of a division between persons and things. No other principle is so deeply rooted in our perception and in our moral conscience..."

What is important is not this difference, but how this differentiation comes to be decided and justified. In order for something to have anything like moral or legal status, it would need to be recognized as another subject and not just an object, for example, a tool or an instrumental means. Standard approaches to addressing and resolving these matters typically proceed, as Coeckelbergh (2012: 14) points out, by following a rather simple and straight-forward decision-making process or what could be called a moral status algorithm:

1) Having property P is sufficient for moral status S
2) Entity E has property P
3) Entity E has moral status S

In this transaction, *we* (and who is included in this first-person plural pronoun is not without consequences) first make a determination as to what ontological property or set of properties are sufficient for something to have a particular claim to moral recognition and respect. In effect, there needs to be a prior identification of what are determined to be the essential qualifying criteria that are needed for "something" to be recognized as "someone" (Spaeman 2006). We then investigate whether an entity, i.e. a robot or AI device, either currently existing or theoretically possible, actually possesses that property or set of properties (or not). Finally, and by applying the criteria decided in step one to the entity identified in step two, it is possible to "objectively" determine whether the entity in question either can or cannot have a claim to moral status or is to be regarded as a mere thing.

This way of proceeding sounds intuitively correct and natural. On this account, questions regarding moral status—decisive questions that decide where to draw the line separating things from persons—are firmly anchored in and justified by the essential nature or being of the entity that is determined to possess them. In this transaction, what something *is* determines how it *ought* to be treated. Or to put it in more formalistic terminology: ontology precedes and determines social, moral, and even legal status. But there are three problems with the approach. The first two—determination and definition—concern complications with the major premise; the third—detection—concerns problems affecting the minor premise.

2.1 Determination

How does one determine which exact property or set of properties are necessary and sufficient for something to have moral status or, as Hannah Arendt (1968: 296) puts it, "the right to have rights?" In other words, which one, or ones, count? The history of moral philosophy can, in fact, be read as something of an on-going debate and struggle over this matter with different properties vying for attention at different times. And in this process many properties—that at one time seemed both necessary and sufficient—have turned out to be either spurious, prejudicial, or both.

Take for example a rather brutal action recalled by the naturalist Aldo Leopold (1966: 237) at the beginning of his seminal essay on environmental ethics: "When god-like Odysseus, returned from the wars in Troy, he hanged all on one rope a dozen slave-girls of his household whom he suspected of misbehavior during his absence. This hanging involved no question of propriety. The girls were property. The disposal of property was then, as now, a matter of expediency, not of right and wrong." At the time Odysseus is reported to have done this, only male heads of the household were considered legitimate moral and legal subjects. Everything else—his women, his children, his animals—were property that could be disposed of without any ethical worries or reflection whatsoever. But from where we stand now, the property "male head of the household" is clearly a spurious and prejudicial criterion for determining moral status.

Similar problems are encountered with, for example, the property of rationality, which is the criterion that eventually replaces the seemingly spurious "male head of the household." When Kant (1985: 17) had defined morality as involving the rational determination of the will, non-human animals, which do not (at least since Descartes had decided that animals were mindless mechanisms) possess reason, are immediately and categorically excluded from consideration. It is because the human being possesses reason, that he—and "human being," in this case, was principally defined and characterized as male, which was the "oversight" Mary Wollstonecraft sought to address by way of her *Vindication of the Rights of Women*—is raised above the instinctual behavior of a mere brutes and able to act according to the principles of pure practical reason (Kant 1985: 63).

The property of reason, however, is contested by efforts in animal rights philosophy, which begins, according to Peter Singer, with a critical response issued by Jeremy Bentham (2005: 283): "The question is not, 'Can they reason?' nor, 'Can they talk?' but 'Can they suffer?'" For Singer, the morally relevant property is not speech nor reason, which he believes sets the bar for moral inclusion too high, but sentience and the capability to suffer. In the book *Animal Liberation* (1975) and subsequent writings, Singer argues that any sentient entity, and thus any being that can suffer, has an interest in not suffering and therefore deserves to have that interest taken into account. Tom Regan, however, disputes this determination, and focuses his "animal rights" thinking on an entirely different property. According to Regan, the morally significant property is not rationality or sentience but what he calls "subject-of-a-life" (Regan 1983: 243). Following this determination, Regan argues that many animals, but not all animals (and this qualification is important, because the vast majority of animal are actually excluded from his brand of "animal rights"), are "subjects-of-a-life": they have wants, preferences, beliefs, feelings, etc. and their welfare matters to them (Regan 1983). Although these two formulations of animal rights effectively challenge the anthropocentric tradition in moral philosophy, there remains disagreements about which exact property is the necessary and sufficient condition for moral consideration.

2.2 Definition

Irrespective of which property (or set of properties) is selected, they each have problems with definition. Take, for example, the property of consciousness, which is often utilized in the discussions and debates regarding moral status for intelligent machines and artifacts. Unfortunately, there is no univocal and widely accepted definition. The problem, as Max Velmans (2000: 5) points out, is that the term unfortunately "means many different things to many different people, and no universally agreed core meaning exists." In fact, if there is any general agreement among philosophers, psychologists, cognitive scientists, neurobiologists, AI researchers, and robotics engineers regarding the property of consciousness, it is that there is little or no agreement when it comes to defining and characterizing the concept. Although consciousness, as the theologian Anne Foerst remarks, is the secular and supposedly more "scientific" replacement for the occultish "soul", it turns out to be just as much an occult property (Benford/Malartre 2007: 162).

Other essential properties do not do much better. Suffering and the experience of pain—which is the property usually deployed in non-standard patient-oriented approaches like animal rights philosophy—is just as problematic, as Daniel Dennett demonstrates in "Why You Cannot Make a Computer that Feels Pain." In this provocatively titled essay, Dennett imagines trying to disprove the standard argument for human (and animal) exceptionalism "by actually writing a pain program, or designing a pain-feeling robot" (Dennett 1998: 191). At the end of what turns out to be a rather protracted and detailed consideration of the problem, Dennett concludes that we cannot, in fact, make a computer that feels pain. But the reason for drawing this conclusion does not derive from what one might expect. According to Dennett, the reason you cannot make a computer that feels pain is not the result of some technological limitation with the mechanism or its programming. It is due to the fact that we remain unable to decide what pain is in the first place. What Dennett demonstrates, therefore, is not that some workable concept of pain cannot come to be instantiated in the mechanism of a computer or a robot, either now or in the foreseeable future, but that the very concept of pain that would be instantiated is already arbitrary, inconclusive, and lacking a clear definition.

2.3 Detection

Most (if not all) of the properties that are considered morally relevant, like the experience of pain or other emotions, are internal mental states or capabilities that are not immediately accessible or directly observable. As Janina Loh (2021: 109) points out, this epistemic uncertainty is not something that is limited to AI and robots: "Actually, we do not only not know what it is like to be a machine or an animal, e.g. a bat—to quote the title of a famous paper by Thomas Nagel (1974). But we also don't really know what it is like to be another human. Because it cannot be determined with unambiguity whether humans are actually equipped with freedom of will and similar abilities. We cannot clearly prove them empirically."

This epistemological barrier is what philosophers commonly call "the problem of other minds." But it is not just a problem for philosophers, as the Brazilian anthropologist Eduardo Viveiros de Castro (2017: 52) explains: "The theological problem of the soul of others became the philosophical puzzle of 'the problem of other minds,' which currently extends so far as to include neurotechnological inquiries on human consciousness, the minds of animals, the intelligence of machines." Although this problem is not necessarily intractable, as Steve Torrance (2014) and others have argued, the fact of the matter is we cannot, as Donna Haraway (2008: 226) characterizes it, ”climb into the heads of others to get the full story from the inside." Even advanced neuroimaging technology like functional magnetic resonance imaging (fMRI) does not provide an easy resolution to this epistemological uncertainty. "This type of technology," as Fabio Tollon (2021: 153) explains, "allows us to peer into the 'moving parts' in the brain which may be correlated with sentience. However, talk of internal states and the talk of how we describe, scientifically, the information that an fMRI machine represents to us are two very different language games."

Responses to this problem, then, typically rely on and mobilize behavioral demonstrations like that devised by Alan Turing for his game of imitation, which inferred machine cognition (an internal state-of-mind) from a demonstration of convincing conversational behavior (an external performance). Even if the behaviors are reasonably convincing, it is still a matter of inferring an internal cause from apparent external behaviors. "We are," as John Basl and Joseph Bowen (2020: 298) explain, "in an epistemologically poor place when it comes to determining what the preferences of an AI are, or what makes it suffer, what it may enjoy, and so on, even if we imagine that the AI is telling us what it 'likes, enjoys, desires, etc.' and behaves accordingly. This is because whatever evidence these behaviors generate is screened off by the fact that the AI might be programmed to behave that way. Yes, the AI convincingly emotes, but it also might have been designed specifically to trick us into thinking it has mental states and emotes because of that despite having no such states."

Consequently, "there is," as Dennett (1998: 172) concludes, "no proving that something that seems to have an inner life does in fact have one." Although philosophers, psychologists, cognitive scientists, and neuroscientists throw an impressive amount of argumentative and experimental effort at the problem, so far it has not been resolvable in any way approaching what would pass for definitive evidence, strictly speaking. In other words, no matter what property is identified, it is always possible to seed reasonable doubt concerning its actual presence. If an AI or a robot, for example, appears to be conscious and therefore a subject of moral concern, all that is necessary to disarm this inference is to point out that it is at least possible that what appears to be conscious behavior is in fact just an effect of clever programming. Likewise, if one seeks to exclude AI or robots from moral consideration on the grounds that they are just things that do not possess real cognitive capabilities, all that is necessary to counter this assertion is to point out that this statement might be true for current systems but may not hold for future systems that have the potential for (and even a high probability of) achieving these threshold conditions. Even if the problem of other minds is not the intractable philosophical dilemma that is often advertised, it is sufficient for sowing doubt about the presence or absence of the qualifying criteria and, by extension, rending decisions about moral status tentative, indeterminate, and uncertain.

3 The Relational Turn

In response to these problems, philosophers—especially in the continental and feminist STS traditions—have advanced other methods for resolving the question of moral status that can be characterized as a relational turn in ethics. This alternative has (at least) three pivotal characteristics:

3.1 Relational

Moral status is decided and conferred not on the basis of subjective or internal properties determined in advance but according to objectively observable, extrinsic social relationships. "Moral consideration," as Coeckelbergh (2010: 214) describes it, "is no longer seen as being 'intrinsic' to the entity: instead it is seen as something that is 'extrinsic': it is attributed to entities within social relations and within a social context." As we encounter and interact with others—whether they be another human person, a non-human animal, or a seemingly intelligent machine—it is first and foremost experienced in relation to us. Consequently, the question of moral status does not depend on what the other is in its essence but on how she/he/it (and pronouns matter here) stands in relationship to us and how we decide, in the face of the other (to use Levinasian terminology), to respond. In this transaction "relations are prior to the things related" (Callicott 1989: 110) such that, as Karen Barad (2007: 136-7) has argued, the relationship comes first—in both temporal sequence and status—and takes precedence over the individual relata.

This shift in perspective—a shift that inverts the standard operating procedure by putting the ethical relationship before determinations of ontological conditions—is not just a theoretical proposal; it has, in fact, been experimentally confirmed in numerous social science investigations. The computer as social actor (CASA) studies undertaken by Byron Reeves and Clifford Nass (1997), for example, demonstrated that human users will accord computers and other technological artifacts social standing similar to that of another human person and that this occurs as a product of the extrinsic social interaction, irrespective of the intrinsic properties (actually known or not) of the individual entities involved. Social standing, in other words, is a mindless operation. And these results have been verified in "robot abuse studies," where HRI (human robot interaction) researchers have found that human subjects respond emotionally to robots and express empathic concern for the machines irrespective of the cognitive properties or inner workings of the device.

3.2 Phenomenological

This alternative is phenomenological or (if you prefer) radically empirical in its epistemological commitments. Because moral status is dependent upon extrinsic social circumstances and not internal properties[1], the seemingly irreducible problem of other minds is not some fundamental epistemological limitation that must be addressed and resolved prior to decision making. Instead of being derailed by the epistemological problems and complications of other minds, the relational turn immediately affirms and acknowledges this difficulty as the basic condition of possibility for ethics as such. Consequently, "the ethical relationship," as Emmanuel Levinas (1987: 56) writes, "is not grafted on to an antecedent relationship of cognition; it is a foundation and not a superstructure…It is then more cognitive than cognition itself, and all objectivity must participate in it." It is for this reason that Levinasian philosophy focuses attention not on other minds, but on the *face* of the other. Or as Richard Cohen (2001: 336) succinctly explains in what could be an advertising slogan for Levinasian thought: "Not other 'minds,' mind you, but the 'face' of the other, and the faces of all others."[2]

This also means that the order of precedence in moral decision making should be reversed. Internal properties do not come first and then moral respect follows from this ontological fact. We have things backwards. We project the morally relevant properties onto or into those others who we have already decided to treat as being socially and morally significant. In social situations, then, we always and already decide between *who* counts as morally significant and *what* does not and then retroactively justify these actions by "finding" the essential properties that we believe motivated this decision-making in the first place. Properties, therefore, are not the intrinsic *a priori* condition of possibility for moral status. They are *a posteriori* products of extrinsic social interactions with and in the face of others.

[1] The only property that would be necessary for something (like a rock) to be in relation to something else (like me) is the minimal ontological condition of being.

[2] "Face" in Levinas is not a substantive property that is possessed by an entity. It is (or takes place as) an act or event of "facing." For more on this and its significance for interpretations and applications of Levinasian philosophy, see Silvia Benso's *The Face of Things* (2000).

3.3 Diverse

Finally, making moral status dependent on consciousness or other cognitive capabilities belonging to the individual is thoroughly Cartesian. Other cultures, distributed across time and space, do not divide-up and make sense of the diversity of being in this arguably binary fashion. They perform decisive cuts separating the *who* from the *what* according to other ways of seeing, valuing, and acting. Following the insights of Josh Gellers (2020), we can identify alternative ways of organizing social relationships by considering cosmologies that are not part of the Western philosophical lineage. As Archer Pechawis explains in his contribution to the essay "Making Kin with Machines":

> *nēhiyawēwin* (the Plains Cree language) divides everything into two primary categories: animate and inanimate. One is not 'better' than the other, they are merely different states of being. These categories are flexible: certain toys are inanimate until a child is playing with them, during which time they are animate. A record player is considered animate while a record, radio, or television set is inanimate. But animate or inanimate, all things have a place in our circle of kinship or *wahkohtowin* (Lewis et al. 2018).

This alternative formulation runs counter to the dominant ways of thinking, seeing the boundary between what Western ontologies call "person" and "thing" as being endlessly flexible, permeable, and more of a continuum than an exclusive opposition.

Similar opportunities/challenges are available by way of other non-Western religious and philosophical traditions. In her investigation of the social position of robots in Japan, Jennifer Robertson (2014: 576) finds a remarkably different way of organizing the difference between living persons and artificially designed/manufactured things:

> "*Inochi*, the Japanese word for 'life,' encompasses three basic, seemingly contradictory but inter-articulated meanings: a power that infuses sentient beings from generation to generation; a period between birth and death; and, most relevant to robots, the most essential quality of something, whether organic (natural) or manufactured. Thus robots are experienced as 'living' things. The important point to remember here is that there is no ontological pressure to make distinctions between organic/ inorganic, animate/inanimate, human/nonhuman forms. On the contrary, all of these forms are linked to form a continuous network of beings."

These are not the only available alternatives, and, by citing these two instances, the intention is not to suggest that these different ways of thinking difference differently are somehow "better" than those developed in Western philosophical and religious traditions. In fact (and this is where things get really complicated), making and operating on that kind of assumption would itself be an instance of "orientalism" (Said 1994), which always sought new resources derived from exoticized Others in order to rehabilitate and ensure the continued success of Western hegemony. The alternatives, by contrast, are just different and, in being different, offer the opportunity for critically questioning what is assumed to be true and often goes by without saying. Gesturing in the direction of other ways of thinking and being can have the effect of shaking one's often unquestioned confidence in cultural constructs that are already not natural, universal, nor eternally true.

4 Objections and Replies

The relational turn introduces an alternative that supplies other ways of responding to and taking responsibility for others and other forms of otherness. But it is by no means a panacea or some kind of moral theory of everything. It just arranges for other kinds of questions and modes of inquiry that are seemingly more attentive to the exigencies of life as it is encountered here and now at the beginning of the 21st century. Having said that, it is important to recognize that relational ethics is not without its own set of unique challenges—three in particular.

4.1 Relativism

For all its opportunities, the relational turn risks exposure to the charge of moral relativism, or as Charles Ess (1996: 204) explains "the claim that no universally valid beliefs or values exist." To put it rather bluntly, if moral status is "relational" and open to different decisions concerning others made at different times for different reasons, are we not at risk of affirming an extreme form of moral relativism? Versions of this objection have been brought by a number of critics, including Vincent Müller (2021) and Kęstutis Mosakas (2021). In fact, Mosakas has provided rather extensive diagnosis of the perceived problem in his contribution to John-Stewart Gordon's *Smart Technologies and Fundamental Rights*:

> As Simon Kirchin explains, "the key relativistic thought is that the something that acts as a standard will be different for different people, and that all such standards are equally authoritative" (Kirchin 2012: 15). Particularly problematic is the extreme version, which denies there being any moral judgments or standards that could be objectively true or false (in contrast to moderate versions that do admit of a certain degree of objectivity) (Moser and Carson 2001: 3). Given the apparent rejection of any such standard by Coeckelbergh and Gunkel, they seem to be hard-pressed to explain how the radically relational ethics (to use Coeckelbergh's own term (Coeckelbergh 2010: 218)) that they are advocating avoids the extreme version (Mosakas 2021: 95).

The perceived problem with relativism (especially the extreme version of it) is that it encourages and supports a situation where anything goes and all things are permitted. But as both Coeckelbergh and I have argued in other contexts (Gunkel 2018 and Coeckelbergh 2020), this particular understanding of "relative" is limited and the product of a culturally specific understanding of and expectation for ethics.

Robert Scott (1967), for instance, understands "relativism" entirely otherwise—as a positive rather than negative term: "Relativism, supposedly, means a standardless society, or at least a maze of differing standards, and thus a cacophony of disparate, and likely selfish, interests. Rather than a standardless society, which is the same as saying no society at all, relativism indicates circumstances in which standards have to be established cooperatively and renewed repeatedly" (Scott 1967: 264). Chares Ess (2009: 21) calls this alternative "ethical pluralism." "Pluralism stands as a third possibility—one that is something of a middle ground between absolutism and relativism… Ethical pluralism requires us to think in a 'both/and' sort of way, as it conjoins both shared norms and their diverse interpretations and applications in different cultures, times, and places." Likewise, Luciano Floridi (2013: 32) advocates a "pluralism without endorsing relativism," calling this third alternative or "middle ground" relationalism.

Others, like Rosi Braidotti, call upon and mobilize "a form of non-Western perspectivism," which exceeds the grasp of Western epistemology. "Perspectivism," as Viveiro de Castro (2015: 24) explains in his work with Amerindian traditions, "is not relativism, that is the affirmation of the relativity of truth, but relationalism, through which one can affirm *the truth of the relative is the relation*.'" For Braidotti (2019: 90) perspectivism is not just different from but is "the antidote to relativism." "This methodology," as she explains, "respects different viewpoints from equally materially embedded and embodied locations that express the degree of power and quality of experience of different subjects." Braidotti therefore recognizes that what is called "truth" is always formulated and operationalized from a particular subject position, which is dynamic, different, and diverse. The task is not to escape from these differences in order to occupy some fantastic transcendental vantage point but to learn how to take responsibility for these inescapable alterations in perspective and their diverse social, moral, and material consequences. The relational turn, therefore, does not endorse relativism (as it is typically defined) but embodies and operationalizes an ethical pluralism, relationalism, or perspectivism that complicates the simple binary logic that defines relativism in opposition to moral absolutism.

4.2 Dehumanization

If moral status is not substantiated by ontological properties but is the product of external relations with others, doesn't this run the risk, as noted by Anne Gerdes (2015: 274), that we might lose something valuable, that "our human-human relations may be obscured by human-robot relations?" This is precisely the concern of Kathleen Richardson (2019: 1), who argues that the relational turn is just as likely to be (mis)used to instrumentalize and reify human subjects: "But if the machine can become another, what does it say for how robotic and AI scientists conceptualise 'relationship'? Is relationship instrumental? Is relationship mutual and reciprocal?" What worries Richardson is that relational ethics—with its focus on social embedding and externalities—would allow for and justify treating other humans as things and not persons. And her response to and fix for this potential hazard is to retreat to a dogmatic reassertion of human exceptionalism. "Humans are never tools or instruments, even if relations between people take on a formal character…In every encounter we meet each other as persons, members of a common humanity" (Richardson 2019: 1).

A different way to respond to this challenge is to recognize, as Anne Foerst suggests, that otherness is not (not in actual practices, at least) unlimited or absolute: "Each of us only assigns personhood to a very few people. The ethical stance is always that we have to assign personhood to everyone, but in reality we don't. We don't care about a million people dying in China of an earthquake, ultimately, in an emotional way. We try to, but we can't really, because we don't share the same physical space. It might be much more important for us if our dog is sick" (Benford/Malartre 2007: 163). This statement is perhaps more honest about the way that moral decision-making and the occurrence of otherness actually transpires. Instead of declaring an absolutist claim to a kind of dogmatic totality, it remains open to particular configurations of otherness that is mobile, flexible, and context-dependent. It is a kind of posthumanist ethic that, as Barad (2007: 136) describes it "doesn't presume the separateness of any-'thing,' let alone the alleged spatial, ontological, and epistemological distinctions that set humans apart."

But there remains, as Gerdes (2015) insightfully recognizes, something in this formulation that is seemingly abrasive to our moral intuitions. This may be due to the fact that this way of thinking does not make a singular and absolute decision about otherness that stands once and for all, such that there is one determination concerning others that decides everything for all time. The encounter with others—the occurrence of face in the face of the other—is something that happens in time and needs to be negotiated and renegotiated. This means that the work of ethics is ostensibly inexhaustible. It is an ongoing and interminable responsibility requiring that one respond and take responsibility for how one responds. Is this way of thinking and doing ethics without risk? Not at all. But the risk is itself the site of ethics and the challenge that one must face in all interactions with others, whether human, animal, or otherwise (cf. Gunkel 2012).

4.3 Performative Contradiction

For all its rhetorical posturing, the relational turn still seems to be inescapably anthropocentric and dependent on properties. This objection is something that is developed by Henrik Skaug Sætra in the essay "Challenging the Neo-Anthropocentric Relational Approach to Robot Rights." Sætra directs his critical efforts to what he identifies as two performative contradictions, where what is espoused by the relational turn betrays or appears to be inconsistent with what it actually does. "Relationalism," as he (2021: 1) explains, "purportedly opens the door for considering robot rights and moving past anthropocentrism. However, I argue that relationalism is, quite to the contrary, a form of neo-anthropocentrism that re-centers human beings and their unique ontological properties, perceptions, and values."

The first critical target in this effort is anthropocentrism or (better stated) its opposite. According to Sætra's reading of the literature, the relational turn promotes itself as being non-anthropocentric but actually is anthropocentric in practice. "My first objection is that relationalism is arguably deeply anthropocentric because moral standing is derived exclusively from how human beings perceive and form relations with other entities. As we have seen, moral standing is here derived from how something is treated, and not what it is. This means that humans are key to determining value, as it is how entities are treated and perceived by humans that determine their moral standing" (Sætra 2021: 6). This criticism is correct. Or, better stated, it is not at all a criticism but an accurate description and characterization.

First, "humans are key to determining value" insofar as morality is a human endeavor. Faulting relationalism for being organized and directed by concerns about human interests and values would be like faulting physics or chemistry for not developing a purely objective science that is not at all involved with human perceptions, concepts, and modes of understanding. The relational turn, like all forms of what Donna Haraway (1991) calls "situated knowledge," comes from somewhere and is embedded and embodied in a specific subject position. To expect that any form of human knowledge would be able to escape from these human-all-too-human conditions of possibility and operate from some super-human position of transcendental objectivity is a metaphysical fantasy that is reserved for the gods. In other words, the axiological purity that Sætra operationalizes as a kind of "litmus test" is a metaphysical fantasy. So yes, the relational turn, like all moral theories and practices, is dependent upon human capabilities, perceptions, and values. And, like all sciences, the critical task is not and cannot be to escape from these existential preconditions but to learn how to respond to and to take responsibility for them.

Second, Sætra is right to conclude from this that relationalism is *not* non-anthropocentric. But he is incorrect in concluding that this double negative implies a positive, namely that it is morally anthropocentric. Non-anthropocentric ethical theories, as Sætra characterizes and explains, include a number of moral innovations that aim to decenter human exceptionalism and culminate, for him at least, in ethical biocentrism: "As compared to the previous type of non-anthropocentrism, ethical biocentrism does not require us to uncover, or conjure up, the interests, preferences, etc., of other entities. Instead, they are considered valuable just because of being what they are, which is why the terms intrinsic or inherent value are often used" (Sætra 2021: 5). But this is not—at least in terms of logical structure—really all that different. Like the anthropocentric model that it contests, ethical biocentrism is still an ontology-driven transaction, where what something is—its being what it is—determines intrinsic or inherent value. The problem, then, is not just with anthropocentrism or its alternatives, but with any and all "epistemic centrisms" which, as Loh (2021: 109) points out "remain committed to the paternalism implicit in the subject-object dichotomy."

The relational turn does not play by these rules. It deliberately flips the script on this entire metaphysical transaction. Following the innovations of Levinas (1969: 304), who famously overturned 2000+ years of Western philosophy by proclaiming that ethics is first philosophy, the relational turn puts the moral relationship *first* in terms of both sequence and status. Or as Barad (2007: 139) describes it, the primary unit "is not independent objects with inherent boundaries and properties," but relations—"relations without preexisting relata."[3] This fundamental change in perspective produces something outside the orbit of either anthropocentrism or its non-anthropocentric others, producing an "eccentric moral theory" (Gunkel 2018: 164) that deconstructs the very difference that distinguishes anthropocentrism from its various alternatives.

This leads to the second critical target, which concerns the status and function of properties. "My second objection," Sætra (2021: 7) writes, "is that relationalism is in reality a camouflaged variety of the properties-based approach. This is so because how we relate to other entities is determined by the properties of these others." In other words, the relational turn can say that it puts relations before relata and makes determinations about moral status dependent on "how something is treated, and not what it is" (Sætra 2021: 6). But this is just patently false, because properties still matter. "How we relate to someone, and how an entity acts, is dependent on their properties." Again Sætra right, but not for the right reasons.

Properties do play a role in moral decision making, and they can be a useful and expedient heuristic for responding to and taking responsibility in the face of others. What is at issue, then, is not their importance but their status and function. As Gellers points out, properties are not antithetical to or excluded from relationalism, they are just recontextualized and understood

[3] This does not mean that all things are essentially nothing outside of being relata. The thing-in-itself (to use Kantian terminology) is ontologically consistent in and of itself. The thing as it stands in relationship to another—as relata—is dependent upon the terms and conditions of the relationship. That fact does not mean (continuing with the Kantian formulation) that there is no *Ding an sich*. It is not nothing; it is just epistemologically inaccessible as it is in itself.

in relational terms. "Coeckelbergh," Gellers (2020: 19) writes, "does not foreclose the possibility that properties may play a role in a relational approach to moral consideration. Instead, he leaves room for "properties-as-they-appear-to-us within a social-relational, social-ecological context (Coeckelbergh 2010: 219)." In other words, the properties that are determined to belong to an entity are actually a phenomenal effect of the relationship and not an antecedent ontological condition and cause. This flips the script on things.

In moral philosophy—at least its standard Western varieties—what something is commonly determines how it ought to be treated. Or as Luciano Floridi (2013: 116) describes it: "what the entity is determines the degree of moral value it enjoys, if any." According to this largely unchallenged standard operating procedure, the question concerning the status of others—whether they are someone who matters or something that does not—is entirely dependent on and derived from what it is and what capabilities it possesses. Ontology, therefore, is *first* in both procedural sequence and status. Sætra not only endorses this way of thinking but normalizes and naturalizes it, even though it is the product of a specific philosophical tradition and culture.

The relational turn not only challenges this way of thinking but deliberately reverses its procedure. This does not diminish the role of properties, it simply inverts the direction of the derivation. The morally significant properties—those ontological criteria that we had assumed grounded moral respect—are actually what Slavoj Žižek (2008: 209) calls "retroactively (presup)posited" as the result of and as justification for prior decisions made in the face of social interactions with others. Consequently, even before we know anything at all about what something is in its essence, we have already been called upon and obligated to make a decisive response.[4]

To give it a Kantian spin, we can say that what something is in itself—*das Ding an sich*—is forever inaccessible insofar as all we ever have access to is how something appears to be relative to us. Whatever we think it is in-itself is the result of something we project onto or into it after the fact. So it is not accurate to conclude that "relationalism is in reality a camouflaged variety of the properties-based approach." Such a conclusion is possible if and only if one normalizes and naturalizes the standard derivation of "ought" from "is." It is just as likely—and maybe even more epistemologically honest—to conclude

[4] For this reason, relations are neither an ontological criterion nor an epistemic category. They are the prior ethical condition. This is why, for Levinas (1969, 304), it is ethics, and not ontology or epistemology, that is "first philosophy."

that what is actually an effect of embedded and embodied interactions with others has been mistakenly dressed-up and masquerading as a cause.

5 Conclusions

The question concerning the moral status of AI and robots is not really about the artifact. It is about us and the limits of who is included in and what comes to be excluded from that first-person plural pronoun, "we." It is about how *we* decide—together and across differences—to respond to and take responsibility for our shared social reality. It is, then, in responding to the moral opportunities and challenges posed by seemingly intelligent and social artifacts that we are called to take responsibility for ourselves, for our world, and for those others who are encountered here.

In devising responses to these challenges, we can obviously deploy the standard properties approach. This method has the weight of history behind it and therefore constitutes what can be called the default setting for addressing questions concerning moral status. It is, to use the terminology of Thomas Kuhn (1996), widely accepted as "normal science." But this normalized approach, for all its advantages, also has demonstrated difficulties with the determination, definition, and detection of the qualifying essential properties. This does not mean, it is important to point out, that the properties approach is somehow wrong, misguided, or refuted on this account. It just means that this way of thinking—despite its almost unquestioned acceptance as normal science within Western traditions—has limitations and that these limitations are becoming increasingly evident in the face or the faceplate of AI and robots—in the face of others who are and remain otherwise.

To put it in terms borrowed from Žižek (2006), the properties approach, although appearing to be the right place to begin thinking about and resolving the question of machine moral standing, may turn out to be the "wrong question" and even an obstacle to its solution. As an alternative, the relational turn formulates an approach to addressing the question of moral status that is situated and oriented otherwise. This alternative circumvents many of the problems encountered in the properties approach by arranging for an ethics that is relational, phenomenological, and diverse. Whether this alternative provides for a better way to formulate moral decisions is something that will need to be determined and decided in the face of others.

References

Arendt, Hannah (1968): The Origins of Totalitarianism, New York: Harcourt Books.

Barad, Karen (2007): Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning, Durham, NC: Duke University Press.

Basl, John/Bowen, Joseph (2020): "AI as a Moral Right-Holder." In: Markus D. Dubber/Frank Pasquale/Sunit Das (eds.), The Oxford Handbook of Ethics of AI, New York: Oxford University Press, pp. 289-306.

Benford, Gregory/Malartre, Elisabeth (2007): Beyond Human: Living with Robots and Cyborgs, New York: Tom Doherty.

Benso, Silva (2000): The Face of Things: A Different Side of Ethics. Albany, NY: SUNY Press.

Bentham, Jeremy (2005 [1789]): An Introduction to the Principles of Morals and Legislation. New York: Oxford University Press.

Braidotti, Rosi (2019): Posthuman Knowledge, Cambridge: Polity Press.

Callicott, J. Baird (1989): In Defense of the Land Ethic: Essays in Environmental Philosophy, Albany, NY: State University of New York Press.

Coeckelbergh, Mark (2010): "Robot Rights? Towards a Social-Relational Justification of Moral Consideration." In: Ethics and Information Technology 12/3, pp. 209-221.

Coeckelbergh, Mark (2012): Growing Moral Relations: Critique of Moral Status Ascription, New York: Palgrave Macmillan.

Cohen, Richard (2001): Ethics, Exegesis, and Philosophy: Interpretation After Levinas, Cambridge: Cambridge University Press.

Dennett, Daniel (1998): Mindstorms, Cambridge, MA: MIT Press.

Esposito, Roberto (2015): Persons and Things: From the Body's Point of View. Translated by Zakiya Hanafi, Cambridge: Polity Press.

Ess, Charles (1996): "The Political Computer: Democracy, CMC, and Habermas." In: Charles Ess (ed.), Philosophical Perspectives on Computer-Mediated Communication, Albany, NY: SUNY Press, pp. 196-230.

Ess, Charles (2009): Digital Media Ethics, Cambridge: Polity Press.

Floridi, Luciano (2013): The Ethics of Information, New York: Oxford University Press.

Gellers, Joshua C. (2020): The Rights of Robots: Artificial Intelligence, Animal and Environmental Law, New York: Routledge.

Gerdes, Anne (2015): “The Issue of Moral Consideration in Robot Ethics.” In: ACM SIGCAS Computers & Society 45/3, pp. 274-280.

Gunkel, David J. (2012): The Machine Question: Critical Perspectives on AI, Robots and Ethics. Cambridge, MA: MIT Press.

Gunkel, David J. (2018): Robot Rights, Cambridge, MA: MIT Press.

Haraway, Donna (1991): Simians, Cyborgs and Women: The Reinvention of Nature, New York: Routledge.

Haraway, Donna (2008): When Species Meet, Minneapolis, MN: University of Minnesota Press.

Kant, Immanuel (1985 [1788]): Critique of Practical Reason. Translated by Lewis White Beck, New York: Macmillan.

Kant, Immanuel (2012 [1786]): Groundwork of the Metaphysics of Morals. Translated by Mary Gregor and Jens Timmermann, Cambridge: Cambridge University Press.

Kirchin, Simon (2012): Metaethics, New York: Palgrave Macmillan.

Kuhn, Thomas S. (1996 [1962]): The Structure of Scientific Revolutions, Chicago: University of Chicago Press.

Leopold, Aldo (1966): A Sand County Almanac, New York: Oxford University Press.

Levinas, Emmanuel (1969): Totality and Infinity: An Essay on Exteriority. Translated by Alphonso Lingis, Pittsburgh, PA: Duquesne University.

Levinas, Emmanuel (1987): Collected Philosophical Papers. Translated by Alphonso Lingis: Dordrecht: Martinus Nijhoff.

Lewis, Jason Edward/Arista, Noelani/Pechawis, Archer/Kite, Suzanne (2018): “Making Kin with the Machines.” In: Journal of Design and Science 16; https://doi.org/10.21428/bfafd97b.

Loh, Janina (2021): “Ascribing Rights to Robots as Potential Moral Patients.” In: John-Stewart Gordon (ed.), Smart Technologies and Fundamental Rights, Leiden and Boston: Brill Rodopi, pp. 101-126.

Mosakas, Kęstutis (2021): “Machine Moral Standing: In Defense of the Standard Properties-Based View.” In: John-Stewart Gordon (ed.), Smart Technologies and Fundamental Rights, Leiden and Boston: Brill Rodopi, pp. 73-100.

Moser, Paul K./Carson, Thomas L. (2001): Moral Relativism: A Reader, New York: Oxford University Press.

Müller, Vincent C. (2021): “Is it Time for Robot Rights? Moral Status in Artificial Entities.” In: Ethics and Information Technology 23, pp. 579-587.

Nagel, Thomas (1974): "What's it Like to be a Bat." In: The Philosophical Review 83/4, pp. 435-450.

Reeves, Byron/Nass, Clifford (1996): The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places, Cambridge: Cambridge University Press.

Regan, Tom (1983): The Case for Animal Rights, Berkeley, CA: University of California Press.

Richardson, Kathleen (2019): Rethinking the I-You Relation Through Dialogical Philosophy in the Ethics of AI and Robotics. In: AI & Society 34, pp. 1-2.

Robertson, Jennifer (2014): "Human Rights vs. Robot Rights: Forecasts from Japan." In: Critical Asian Studies 46/4, pp. 571-598.

Sætra, Henrik Skaug (2021): "Challenging the Neo-Anthropocentric Relational Approach to Robot Rights." In: Frontiers in Robotics and AI 8, pp. 1-9.

Said, Edward (1994 [1978]). Orientalism, New York: Vintage Books.

Scott, Robert L. (1967): "On Viewing Rhetoric as Epistemic." In: Central States Speech Journal 18, pp. 9-17.

Singer, Peter (1975): Animal Liberation: A New Ethics for Our Treatment of Animals, New York: New York Review of Books.

Spaemann, Robert (2006): Persons: The Difference between "Someone" and "Something." Translated by Oliver O'Donovan, New York: Oxford University Press.

Tollon, Fabio (2021): "The Artificial View: Toward a Non-Anthropocentric Account of Moral Patiency." In: Ethics and Information Technology 23/1, pp. 147-155.

Torrance, Steve (2014): "Artificial Consciousness and Artificial Ethics: Between Realism and Social Relationism." In: Philosophy & Technology 27/1, pp. 9-29.

Velmans, Max (2000): Understanding Consciousness, New York: Routledge.

Viveiros de Castro, Eduardo (2015): The Relative Native: Essays on Indigenous Conceptual Worlds. Translated by Martin Holbraad, David Rodgers and Julia Sauma, Chicago: HAU Press.

Viveiros de Castro, Eduardo (2017): Cannibal Metaphysics: For a Post-Structural Anthropology. Translated by Peter Skafish, Minneapolis, MN: University of Minnesota Press.

Wollstonecraft, Mary (2004 [1792]): A Vindication of the Rights of Woman. New York: Penguin Classics.

Žižek, Slavoj (2006): "Philosophy, the 'Unknown Knowns,' and the Public Use of Reason." In: Topoi 25, pp. 137-142.

Žižek, Slavoj (2008): For They Know Not What They Do: Enjoyment as a Political Factor, London: Verso.

From Tool Use to Social Interactions

Anna Strasser

1 Introduction

Human-machine interactions are often exclusively described as tool use and not as social interactions. However, this contradicts experiences we can make in interactions with artificial agents, especially when social robotics come into play. Without a doubt, humans have the ability to bond emotionally not only with living beings but also with inanimate artificial agents. Consequently, we treat artificial agents as if they were social agents (cf. Carpenter 2016; Darling 2016). However, behaving socially in front of artificial agents might not yet fully justify qualifying them as social agents and attributing social agency to them.

This tension provides a motivation to question the widespread restrictive conception of sociality, in particular of social agency, which excludes inanimate objects as potential participants in social interactions. According to this view, it seems widely accepted that being alive constitutes a minimal condition (necessary but not sufficient) to be considered a social agent. Following this line of thinking, it is argued that artificial agents cannot qualify as participants in social interactions because they lack the essential properties of living beings. Particularly in the debates about joint actions in the field of philosophy of mind, it becomes obvious that definitions of socio-cognitive abilities that lay a foundation for the attribution of social agency exclude artificial agents from the outset.[1]

[1] I am not questioning that humans can behave socially toward all sorts of entities. However, I think that this should be separated from the question of whether it is justified to attribute socio-cognitive abilities to these entities. At this point, it should be noted that in other fields of research the notion of 'social interaction' seems to be understood more broadly, namely in the sense that everything is understood as social interaction that follows a specific pattern of human behavior. Thus, all interactions in which we

In fact, one can observe that notions describing agency, the ability to act jointly, or mindreading are restrictive even in a more radical sense (cf. Davidson 1980; Bratman 2014; Fodor 1992). They characterize such abilities as if they were unique to sophisticated human beings only. One factor contributing to restrictive conceptions might be that philosophy's main objective is to create sharp, clear-cut notions, resulting in a general tendency to focus on ideal cases with rather demanding conditions.

Therefore, it is not surprising that some sophisticated terminology of philosophy already reaches its limits when it comes to socio-cognitive abilities of other living agents, such as children or non-human animals (cf. Brownell 2011; Heyes 2014, 2015; Pacherie 2013; Perler 2005; Premack/Woodruff 1978; Vesper et al. 2010; Warneken et al. 2006). In contrast, developmental psychology and animal cognition demonstrate gradual appearances and multiple realizations of socio-cognitive abilities, but these cannot be captured by the aforementioned sophisticated terminology.

Proposals responding to these shortcomings introduce terms such as simple forms, proto-cases, or quasi-states to extend the restrictive conceptual framework to a wider range of cases. For example, Perler and Wild talk of non-human animals as having simple thoughts – "simple 'Geisthaber'" (2005: 70) – to describe the thinking abilities of non-human animals.

So-called minimal approaches follow a comparable strategy, and I aim to show that their strategy offers a promising starting point for characterizing the abilities of artificial agents that cannot be captured by the sophisticated terminology. By challenging overly demanding conditions of certain philosophical notions, minimal approaches can cover a broader range of socio-cognitive abilities. Examples are notions such as *minimal mindreading, minimal sense of commitment*, and *shared intentions lite* (cf. Butterfill/Apperly 2013; Michael et al. 2016; Pacherie 2013).

Also in the debates of moral philosophy, there are more and more positions that question all too demanding conditions regarding (moral) agency. Especially with respect to the evaluation of interactions with potentially social robots, arguments are made for an extension of the conceptual framework (Wallach/Allen 2009). This becomes particularly evident concerning concepts that describe assumed unique human capabilities such as autonomy. Thus,

behave socially towards entities count as social interactions. This also means that artificial systems can trigger social behavior, regardless of whether a social agency can be attributed to them.

in debates about moral agency in general, a more fine-grained differentiation in terms of autonomy is proposed (Darwall 2006: 265). Instead of necessarily presupposing full-fledged autonomy, several degrees of autonomy are conceptualized. Furthermore, the necessity of other demanding conditions is questioned. For example, Floridi and Sanders (2004) claim that consciousness, intentionality, mental states, and intelligence are not necessary for characterizing moral agency of artificial agents. According to them, it is sufficient to require negative autonomy (absence of external force and control), interactivity, and adaptability (see also, Wallach/Allen 2009; Misselhorn 2018).

Reflecting on the increasingly important role artificial agents play in our social life, controversial positions have been developed when assessing the status of artificial agents. Aside from an instrumentalist view, which is based on a demanding understanding of agency and claims that artificial agents can, in principle, only be considered as tools (cf. Johnson 2006), one can now find positions arguing that artificial agents can be considered as social agents (cf. Darling 2016; Hakli/Seibt 2017). Assuming that the impact of human-machine interactions on our social life will increase and that such interactions, to most appearances, will significantly differ from other tool use cases, this paper aims to elaborate under which circumstances we are justified to consider artificial agents as social agents instead of mere tools.

Arguing for the claim that one should broaden the conceptual framework of social agency if one is to arrive at adequate characterizations of all varieties of social interactions, I use the debate about joint actions in philosophy of mind as an example to explore the extent to which one can ascribe a minimal form of agency to artificial agents that do not meet the demanding conditions of full-fledged agency. The suggested conceptual framework allows for fine-grained distinctions between mere behavior and action (Strasser 2006: 172). Since agency alone is not sufficient to qualify as a social agent in joint actions, I also investigate how to conceptualize the necessary socio-cognitive abilities of artificial agents in order to ascribe social agency to them.

The decision to deal with joint actions should not be understood as an equation of joint actions and social interactions. Joint actions are only one form of many forms of social interactions. They are perhaps one of the most demanding forms of social interaction, and this seems to me to be a particular challenge. My concern is to show precisely for sophisticated social interactions to what extent one can argue that artificial agents can also be qualified as possible participants here. Moreover, the debate about joint actions provides a template for rethinking anthropocentric assumptions. In this context, Christian List (2021) points out an interesting parallel regarding group action and artificial agents. According to him, both phenomena can be understood as interactions with non-human, goal-directed agents (collective agency/artificial agency) that can change the social world. And this can be taken as a reason to reconsider some of our anthropocentric moral assumptions.

Although both conceptual issues and psychological factors contribute to a theoretical foundation on which we can argue for the potential sociality of artificial agents, this paper focuses on the conceptual issues. That is, I am less concerned with psychological factors that contribute to the subjective impression that certain interactions with artificial agents are social interactions than I am with elaborating minimal necessary conditions that must be met by artificial systems to be considered proper participants in social interactions. Therefore, I am primarily addressing the conditions that artificial agents must satisfy so that we are justified in ascribing social agency to them. This may help provide a fully comprehensive basis for an analysis of all factors that contribute to transforming human-machine interactions from the mere use of tools into social interactions.

Questioning that all human-machine interactions can be reduced to mere tool use, I point to social phenomena in which artificial agents are not merely involved but play an active role and show that established notions from the philosophy of mind do not provide an adequate conceptual framework for such phenomena (section 2). To conceptualize these phenomena, one must rethink restrictive conceptions of sociality. Thereby, questioning whether biological constraints inhibit a necessity will be of primary importance (section 3). Understanding joint actions as one interesting and challenging case of a social interaction, I suggest a conceptualization of a non-human-centered version of acting jointly that will specify the conditions artificial agents must meet in order to enter the space of social interaction (section 4). Finally, I point to research findings that indicate to what extent proposed conditions are already met by actual artificial agents (section 5).

2 Varieties of human-machine interactions – the discovery of terra incognita

Humans have entertained manifold types of tool use from the Stone Age to the present day, and tools have changed over time. In general, tool use is understood as the instrumental use of objects to achieve certain goals. For example, using a hammer to drive in a nail is a prototypical case. However, not only inanimate objects can be used as tools, but also humans or other living beings can be used as tools. Being a social agent does not prevent you from being used as a tool. Animate and non-animate entities can be used as tools. However, to be considered an active participant in a social interaction, one must qualify as a social agent, whereby I do not presuppose that all social agents must be animate entities. In our time, there are increased actions involving various types of new technological devices, such as the Internet, cell phones, social networks, and last but not least, chatbots and social robots. In fact, artificial intelligence is shaping many of these new technological devices.

The question I am concerned with is whether all these devices should be considered mere tools. Obviously, tools differ concerning their complexity; we can easily distinguish simple tools such as a hammer from more complex tools such as statistical software or social robots. However, in view of technical devices that display some learning abilities, the differences become greater. Normally, tool use implies to some extent that the user is in control of the tool. However, there are tools that are only to a very small extent under our control; such tools show some degree of autonomy or are even able to adapt and learn. This is reflected, for example, in the distinction between so-called *in-the-loop systems, on-the-loop systems,* and *out-of-the-loop systems* (cf. Leveringhaus 2016: 3; Loh 2019: 33). The former are subjects to human control, whereas *out-of-the-loop systems* describe machines in which humans even do not have an intervention option. In-between, there are *on-the-loop systems*, which have some autonomy, but the human still can intervene. On-the-loop systems can decide and act autonomously, but the final decision remains with the human and thus, it is argued, the responsibility. With respect to the different grades of autonomy, some tools already have agent-like properties. This means that there are interesting differences on the table that require finer conceptual differentiation and may raise the question of whether some tools can not only qualify as agents but also as social agents.

The reasons why describing human-machine interactions as mere tool use contradicts the experiences we make with artificial agents are manifold. Apart from the hypothesis that artificial agents may indeed qualify as a new type of social agents, which is this paper's focus, psychological factors also contribute to the impression that certain human-machine interactions are social interactions. For example, interactions with social robots are perceived as social and not as tool use because these products are specifically designed to trigger the human tendency to anthropomorphize (in the sense of a tendency towards behaving socially). Treating entities socially, we behave as if they were social agents. This can be illustrated by our relation to certain artificial toy pets. Darling (2016) reports from a group of participants that were given a cute robotic dinosaur called *Pleo* to interact with. In the end, they were asked to tie up, strike, and *kill* their *Pleo*. However, due to the human tendency to sociality – in that case, an animomorphization – participants refused to *hurt* their robot. The human ability to emotionally connect not only with living creatures but also with inanimate artificial agents plays an important role in many domains, even in the military sphere, as Carpenter (2016) reported. Undoubtedly, we often interact with artificial agents as if they were social agents and not just tools.

At this point, some positions argue that artificial agents (or actually their creators) simply trick humans into attributing properties to artificial agents falsely and thus conclude that all human-machine interactions are, in the end, just tool use, no matter how social such interactions appear (Bryson 2010). To this end, it might be mentioned that anthropomorphism, including animomorphism, is traditionally seen as a bias, a category mistake in psychology (cf. Damiano/Dumouchel 2018; Loh 2019). However, admitting that there are cases of being tricked does not exclude the possibility of genuine social interactions with artificial agents. Given that some human-machine interactions, especially those with social robots, are significantly different from those that constitute mere tool use (e.g., by involving only simple tools, such as laptops or toasters), I argue that it is reasonable to address these differences by re-examining the conditions artificial agents must fulfill in order to display socio-cognitive abilities and, thus, be considered as a new type of social agent.

Of course, observing a person who spends most of her time with a care robot, one can describe many of the involved interactions as tool use. The person uses the robot as an assistant (tool) that gives support and takes care of activities the person is not able to do. Nevertheless, it is at least conceivable that the person may also communicate with such a robot and satisfy social needs. Regarding those interactions, the question arises whether the care robot takes on the role of a social agent here. It is not necessary to go so far and claim that a care robot could replace a human caregiver. However, it seems reasonable to assume that some of those interactions have the potential to lead to experiences which are strikingly similar to those we make with human counterparts, and this might be due to abilities the robot actually has.

These considerations suggest that not all interactions with social robots should be reduced to instrumental, ordinary tool use. When talking about tool use, one usually assumes that tools are rather passive. That is, we do not assume that our tools are capable of acting autonomously, nor do we expect them to adapt to our behavior or even learn new behaviors and respond to social cues. At least in the last decades, we would have been quite irritated if hammers or bicycles would suddenly smile or say something back. Whereas social bots are explicitly developed to be companions that adapt, learn, and communicate. Some of them are able to process and display social cues.

Moreover, the assumption that certain human-machine interactions are in some ways similar to human-human interactions has already found its way into empirical research. Here, it is assumed that the way humans behave in interactions with artificial agents bears at least some resemblance to the way they behave in interactions between humans because people make socialness attributions (cf. Hortensius et al. 2018). Consequently, experimental protocols with artificial agents are used to gain insightful information about humans' social cognitive mechanisms (cf. Wykowska et al. 2016). If there were no similarities at all, such experiments could not provide any information about human-human interactions.

Although psychological factors help characterize specific features of human-machine interactions, it is important not to rely exclusively on first-person attributions when arguing for potential social human-machine interactions. Humans are prone to treat tools or toys as substitutes for social agents in everyday interactions. A subjective feeling is not yet sufficient to justify attributing socio-cognitive abilities. For a justified attribution, I argue, *it must be shown that artificial agents contribute to the interactions utilizing socio-cognitive abilities*. By fulfilling minimal necessary conditions, artificial agents can prove as active participants contributing to social interactions.

If both parties actively shape certain human-machine interactions, the observed interaction can no longer be described as mere tool use. This leads to the question of specifying conditions for artificial agents' socio-cognitive abilities and thereby expanding the current conceptual framework of sociality. Overall, we are not only dealing with a conceptual question concerning the conditions that artificial agents must fulfill but also with an epistemological question, namely when we are entitled to conclude that these conditions are met.

Of course, one could still decide to expand the conceptual framework of tool use (e.g., by introducing social tools) instead of revising the conceptual framework regarding social agency. I suspect, however, that the notion of tool use would turn out to be inadequate to capture precisely social aspects of such interactions, such as the *reciprocal exchange of social information*. Assuming that certain interactions with artificial agents are somehow located in a *terra incognita*, where the terminology from philosophy of mind cannot conceptualize them, I will argue below that these phenomena have enough similarities[2] with social interactions between humans. Thus it seems reasonable to opt for re-examining the conceptual framework of social agency. My proposal for the conceptual clarification of the minimal necessary conditions in conjunction with an investigation of epistemological questions aims to provide arguments for claiming that artificial agents can qualify as active participants, respectively, as social agents, in social interactions under certain circumstances.

3 Overcoming restrictive conceptions of sociality

Nevertheless, extending the conception of sociality is a challenging endeavor because it contradicts both our common sense and current philosophical ideas of sociality we can find in philosophy of mind. So far, many conceptions of sociality are limited to living beings. Although the demarcation between the living and the non-living is not always that obvious, up to now, it seems to be widespread that being alive is taken as a necessary but not sufficient condition for social agency. However, assuming that certain interactions with artificial agents should be considered social interactions and not tool use, it seems reasonable to rethink the conditions for social agency.

Turning to philosophy of mind, there are typically thought to be two key elements of social interaction partners: the capacity for genuine agency and certain social abilities. However, on very demanding conceptions, such as those elaborated by Davidson (1980; 2001), neither artificial systems, nor certain disabled persons, nor infants, nor non-human animals fulfill either of the requirements. According to Davidson, the capacity for agency, thought, language, and interpretation are interrelated, and only linguistically sophisticated creatures can be genuine agents, and hence genuine social agents. According to this view, socio-cognitive abilities are characterized as if they were only present in sophisticated adult humans. At this point, it may be debatable how often even human adults actually encounter such ideal cases in everyday life. In addition, one can question the consistently anthropocentric

[2] Such similarities are not only grounded in our subjective experiences but also in abilities we are justified ascribing to the artificial agent.

character of those concepts. In fact, research indicates that this demanding conception is too demanding. There are multiple realizations of socio-cognitive abilities in various types of agents, such as infants and non-human animals (cf. Brownell 2011; Heyes 2014, 2015; Pacherie 2013; Perler 2005; Premack/Woodruff 1978; Warneken et al. 2006). And different capacities come online at different stages of development (Perner 1991; Tomasello 2008). This evidence supports the common-sense notion that infants are social beings with whom one can interact socially (Vesper et al. 2010), and it might likewise become part of our common sense to consider certain artificial systems as social interaction partners.

What makes me optimistic regarding arguing for an extension of the restrictive conceptual framework in the philosophy of mind is the fact that attitudes toward the status of social agents have proven mutable throughout human history. Here one can point out that, at least as far as the status of women, children, other ethnic groups, and some non-human animals is concerned, an extension of the class of social agents has already arrived in our common sense. This shows that formerly excluded subjects can be considered proper social agents due to social changes.

If we consider the status of slaves, we can even describe a case in which living beings are deprived of their status as social agents and are instead considered as tools without rights[3]. That this should be changed is beyond question. But it vividly illustrates that even the minimal requirement of being alive is not sufficient to maintain social agent status. This is also illustrated by the widespread assumption that one cannot have social interactions with all kinds of non-human animals, e.g., interactions with mosquitoes.

Although it may sound radical to a Western audience, I argue that even non-living beings can be considered social agents. In this context, it is worth mentioning that outside of Western cultural conceptions, e.g., in Shintoism and Animism, there are already conceptions that characterize objects as animate that are considered inanimate from a Western perspective (see Jensen/Blok 2013).

[3] For example, Aristotle described slaves as “animate tools” in “Politics” (1, 1253b15-55b40).

3.1 Towards a broader conception of sociality

Starting from the assertion that the current terminology in philosophy of mind cannot adequately describe certain human-machine interactions, one has to expand the conceptual framework to capture multiple realizations of social agency.

So far, most objections concerning the classification of artificial agents as social agents are based on arguments claiming that non-living entities lack essential abilities which are necessary for social agency (Nida-Rümelin 2022). In this context, for example, the lack of full-fledged intentionality, free will, and emotional states is cited as a reason for excluding artificial agents (Davidson 1980). In the debates about moral agency, the lack of phenomenal consciousness and, in particular, the lack of the ability to suffer is often taken as a reason why artificial agents cannot be moral objects, respectively moral subjects.

However, following the strategy of above mentioned minimal approaches, one can question the necessity of some conditions by demonstrating that multiple realizations of socio-cognitive abilities are conceivable and part of everyday life. For instance, one feature of social interactions that can be elaborated on is that the individuals involved are able to attribute mental states to each other. This ability of mindreading enables them to anticipate, to some extent, the behavior of their interaction partners; if they were not able to do so, social interactions would become immensely difficult, if not impossible. Similar to Davidson's notion of action, one finds very presuppositional notions of mindreading that exclude many types of agents from the outset (Fodor 1992). In response to the restrictive attribution of full-fledged mindreading, Butterfill and Apperly (2013) developed the notion of *minimal mindreading* by questioning the necessity of overly demanding cognitive resources. Thus, they can characterize automated mindreading in adults as well as abilities of children and other animals (cf. Strasser 2012). Since *minimal mindreading* does not require conscious reasoning, this notion is also suitable, as I will argue, to characterize potential abilities of artificial agents.

3.2 Joint action

Since there are manifold social interactions that I cannot consider in-depth, I decided to focus here on a specific subclass, namely joint actions. One might object that acting jointly is one of the most presuppositional form of a social interaction and that it would be more obvious to start with not-so-demanding forms of social interactions. However, I think that if one can show that it is arguable that artificial systems even can qualify as possible partners in joint actions, then the claim that artificial agents can qualify as social agents is on safer ground.

Joint actions are social interactions in which (at least) two agents cooperate and do things together to reach a common goal. According to Bratman, having a shared intention is taken as the essential condition of joint actions among human adults (Bratman 2014: 152). However, the fulfillment of the conditions for having shared intention proves to be demanding. Just having an intention is not enough. In addition, one has to be able to entertain a specific belief state that enables a relation of interdependence and mutual responsiveness between one's own intentions and the others. In short, acting jointly, you need shared goals. All this, it is assumed, requires the ability to have common knowledge, mastery of mental concepts, and sophisticated mentalization skills. Following Bratman, disabled persons, children, and non-human animals are excluded because they lack sophisticated mental concepts and capacities for explicit commitments. However, this conflicts with both our common sense and empirical data (cf. Brownell 2011). Children are understood as socially interacting beings even though they do not fulfill the demanding conditions of a standard notion of joint action (cf. Vesper et al. 2010). For example, playing hide and seek with children is experienced as a proper joint action. Furthermore, research indicates that not only children but also non-human animals successfully engage in joint actions without fulfilling the demanding conditions of Bratman's notion (cf. Warneken et al. 2006). Asking whether artificial agents might be able to participate in a joint action is just one step further.

The strategy of minimal approaches to capture a broader range of socio-cognitive abilities by proposing minimal versions of established notions is, in my view, a promising starting point to overcome restrictive conceptions of sociality. By assuming multiple realizations, I can offer an extension of the restrictive notions in play and present a way how one can also capture socio-cognitive phenomena with respect to artificial agents. If one can establish the idea of multiple realizations, one can question the absoluteness of the demanding conditions put forward. In the best case, I can thus show on what basis a further discussion of these nevertheless strikingly restrictive theoretical approaches continues to make sense. Instead of assuming full-fledged joint actions in which all participants satisfy the same demanding conditions, one can then show that not all the conditions we require in the human case, such as having emotions, turn out to be necessary for artificial agents as well. On the basis of multiple realizations, one can work out a set of minimal necessary conditions that satisfy the requirements we impose on artificial agents in social interactions.

4 A minimal notion of joint action

To develop a minimal notion of joint action, I start with a rough working definition: Every agent acting jointly must be able to act and coordinate. The supposed necessity of particular involved requirements regarding coordination will be investigated step by step.

4.1 Asymmetric actions

Joint actions involving children or non-human animals already hint at the possibility of multiple realizations in which those conditions can be fulfilled. Due to multiple realizations, the distribution of abilities can vary among the participants. This is what I call asymmetric joint actions. For example, mother-child interactions illustrate that the distribution between participants is not necessarily equal.

Assuming that there are also multiple realizations regarding artificial agents, it is only one step further to suppose that besides joint actions with mixed groups consisting of human adults and children, there may as well be joint actions with humans and artificial agents. Describing human-machine interactions as asymmetric joint actions, there is no need to suppose the very same sets of conditions for artificial agents. Returning to my rough definition of joint action, saying that every agent acting jointly must be able to act and coordinate, the conditions of a *minimal joint action* will presuppose multiple realizations of acting and coordinating, respectively *minimal agency* and *minimal coordination*. Focusing on asymmetric joint actions, the developed notion clarifies which conditions should be imposed on the different participants of such joint actions. A task for future research would be to investigate whether joint actions can also occur between two or more artificial agents, i.e., whether there can be social interactions in which no humans are involved.

4.2 Ability to act

Without a doubt, if you are not able to act, you cannot act jointly. According to philosophy of mind, one of the important features of actions is that they are "intentional under some description" (Davidson 1980: essay 3). However, due to a strong sense of intentionality, this implies various further abilities, such as being equipped with consciousness, generating goals, and making free choices. Full-fledged agency, as, for instance, described by Davidson, includes highly demanding conditions. Besides intentionality, consciousness, the ability to generate goals, and the ability to make free choices, it is also required that acting agents have propositional attitudes and a mastery of language. According to such demanding conditions, non-living beings, non-human animals, people with disabilities, and children cannot act. All that they are capable of is producing more or less complex behavior. Criticism concerning the exclusion of children and non-human animals has already been raised. What has to be shown now is that this critique can also concern the exclusion of non-living artificial agents.

Assuming that the abilities of artificial agents in certain human-machine interactions are neither adequately described by mere behavior nor by full-fledged agency, I argue that one should make a finer-grained differentiation concerning the classes of events, such as behavior and action. With a notion of a *minimal action,* one can capture phenomena in-between mere behavior and full-fledged actions (cf. Strasser 2006). On the basis of a minimal notion, one can then argue in a similar way as Wallach and Allen (2009) do with respect to moral agency, that agency should be understood as a gradual concept. Minimal and full-fledged agency could then characterize the extreme cases of a continuum. In line with the rationale of other minimal approaches, the notion of a *minimal action* can question the necessity regarding some conditions that exclude artificial agents from the outset. Assuming that agency can be reached by interpreting proposed conditions in a weaker sense, the necessity of a full-fledged realization of all requirements is questioned. For example, one can question whether *minimal agency* necessarily requires that the generation of goals occurs in the acting entity. Alternatively, a goal can be generated in another system and then transferred to the minimal acting system. If the minimal acting system can recognize and represent goals as goals, it can still act goal-directed. Likewise, being conscious can turn out to be not a necessary condition. This is not to deny that consciousness plays an important role in human cases. But given that there are multiple realizations, consciousness could be a specific property of living beings (a biological constraint) that lacks necessity with respect to artificial agents.

Roughly speaking, *minimal agency* requires that artificial agents are able to perceive, represent, and process the relevant information (including goals, context, etc.) and must have effectors to perform an action. Of course, the development of such a notion requires more specifications, clarifying the extent to which perception, representation, and processing include abilities to adapt and learn (cf. Strasser 2006, 2015). For the sake of argument, let us assume that one can presuppose minimal agency with respect to artificial agents. It may be important to clarify that this minimal notion requires conditions that the minimal actor must actually fulfill, and should not be confused with attribution practices of other actors, such as those described by Dennett's (1987) intentional stance.

Now, when describing asymmetric joint actions, one can specify two different realizations by referring to the minimal and full-fledged notions of agency. Artificial agents can realize the ability to act in a more minimal way that does not require conscious, mental, or emotional states. At the same time, the more demanding conditions of a full-fledged agency can describe the agency of the human counterpart.

4.3 Coordination

Just being able to act is not sufficient to qualify as a participant in a joint action. In addition, the ability to coordinate is an essential prerequisite for joint actions. This ability plays a crucial role for the social dimension of joint actions. With reference to Bratman's notion of a joint action that requires shared intentions, one can argue that the functional role of coordination is to enable shared intentions. Only if agents work together in an organized way, we can talk of joint actions. Investigating minimal necessary conditions artificial agents have to fulfill to coordinate with human counterparts, I elaborate on three important aspects: reciprocal exchange of social information, mindreading, and commitment.

4.3.1 Reciprocal exchanges of social information

Explorations of human social cognition highlight the importance of social signals (cf. Frith/Frith 2007). Working together in an organized way requires reciprocal exchanges of social information. In human-human interactions, we observe an exchange of a wealth of information transferred by language or other expressive behaviors. In addition to verbal agreements, also social cues such as gestures and facial expressions are exchanged. Deficits in interpreting non-verbal behavior can lead to deficits in social interactions (cf. Mundy et al. 1986; Bogart/Tickle-Degnen 2015). Since humans apply social cues in joint actions, it is a necessary requirement for artificial agents to interpret and send social cues back to their human counterparts. Regarding tool use, there is no need for a reciprocal exchange of social information. In contrast, processing social cues seems to be a distinguishing feature of social interactions. Consequently, the first requirement for coordination with human counterparts concerns the ability to handle social cues. I aim to show that it is not necessary that artificial agents actually have emotional and mental states. From the perspective of establishing minimally necessary criteria, I argue that it is sufficient for artificial agents to have the ability to express and interpret social signals. Analogous to the ability to act, it can be argued here that there are multiple realizations of how a condition can be satisfied. For example, instead of requiring emotional or mental states, one could implement functions that are realized by emotional or mental states in the human case. This is in line with Wallach and Allen (2009), who speak of functional equivalence in this context. The general ability to process and interpret social information constitutes social competence, which seems to be an essential prerequisite for any kind of social interaction.[4]

4.3.2 Mindreading

Another relevant aspect of coordination in joint actions consists in the fact that participants normally are able to anticipate to some extent what the other agent will do next. In the humanities and natural sciences, one aspect of such anticipation abilities is discussed under the label *mindreading* or *Theory of Mind* (cf. Fodor 1992; Fletcher/Carruthers 2013; Gopnik 2003, Nichols/ Stich 2003). Once again, there are notions that are tailored to human adults only. However, research indicates that mindreading abilities are present in children and non-human animals. This motivated Butterfill and Apperly (2013) to develop the notion of *minimal mindreading*, which can account for a broader

[4] At this point, one might get the impression that the development of minimal notions does take place in the spirit of Dennett's intentional stance (1987). However, I think there is a difference between requiring conditions to be met by an entity in order to be justified in attributing an ability to that entity and adopting an intentional stance because it is practicable.

range of mindreading. Questioning the necessity of overly demanding cognitive resources, such as representing a full range of complex mental states and a mastery of language, they elaborated minimal necessary conditions that can explain success in mindreading tasks. For example, the full range of representations of complex mental states is replaced by representations of less complex mental states, specified as encounterings and registrations, and it is argued that they are sufficient to anticipate the behavior of other agents in an efficient, automatic, fast, and robust manner (Butterfill/Apperly 2013: 18). Most significantly, with regard to artificial agents, *minimal mindreading* does not require conscious reasoning. Thus, the notion of *minimal mindreading* can account for mindreading abilities, such as automatic mindreading in human adults and mindreading abilities in children and non-human animals. Therefore, the second requirement for coordination with human counterparts concerns the ability to display *minimal mindreading* abilities.

4.3.3 Commitment

Besides processing social information and anticipating the behavior of the counterpart, it is crucial for the success of human-human joint actions that both agents can rely on the contribution of the other agent. This means both parties are committed to sticking to the joint action in the ideal case. Another important function of commitments is that they make agents' behavior easier to predict (Michael/Pacherie 2015: 100). It is not surprising that some notions of a (strict) commitment are tailored to human adults only (cf. Shpall 2014). Strict commitments are characterized as a triadic relation among two agents and an action. Involved agents mutually have certain expectations and motivations concerning this action.

To explore in what sense commitment is important for human-machine interactions, the minimal approach by Michael et al. (2016) offers a good starting point. Michael and colleagues argue that components of a strict commitment can be dissociated. They suggest that a single occurrence of one component can already be treated as a sufficient condition for a *minimal sense of commitment*.

Following the idea that components of a commitment can be disassociated, one can describe a *minimal sense of commitment* on the human side with respect to human-machine interactions. Regardless of whether the artificial agent is committed, a human counterpart can have an expectation of what the artificial agent should do and thereby assume that the artificial agent is committed. Alternatively, the human can be motivated to contribute because she implicitly assumes that the artificial agent is expecting this. By disassociating expectations from corresponding motivations, the *minimal sense of commitment* does not rely on the corresponding counterpart's abilities.

Focusing on the side of the artificial agents, one might object that we are not justified to ascribe motivations or expectations to artificial agents. Therefore, analyzing the side of the artificial agents with respect to a *minimal sense of commitment* is challenging. Avoiding the requirement of mental and emotional states, such as expectation and motivation, I suggest allowing functional corresponding states of expecting and feeling motivated (cp. functional equivalence Wallach & Allen 2009: 68). Thereby, I follow the strategy of questioning whether certain biological constraints are necessary. Formulating conditions of how a *minimal sense of commitment* can be realized in human-machine interactions, the third requirement for artificial agents demands an ability to interpret signs of being committed regarding the human counterpart as well as the ability to react by signaling expectations and motivations in order to contribute to the joint action.

In sum, to qualify as a proper participant in a joint action with a human counterpart, artificial agents have to fulfill conditions ensuring that we can ascribe *minimal agency* and an active contribution to the coordination of this joint action. Regarding coordination, the artificial agent has to be able to join a reciprocal exchange of social information. Furthermore, artificial agents should display *minimal mindreading abilities* as well as the ability to exhibit and elicit a *minimal sense of commitment*.

5 Pieces of the puzzle found in AI

So far, the development of a notion of *minimal joint actions* has remained on a theoretical, conceptual level. Approaching epistemological questions as to whether we are justified in ascribing artificial agents the required abilities, one has to evaluate actual abilities of artificial agents. The following examples show that, though distributed over distinct systems, proposed conditions can be fulfilled in principle. Consequently, it seems conceivable that even a single artificial agent may fulfill all conditions to be a participant in a *minimal joint action* in the near future. Nevertheless, one has to admit that the claim that artificial agents could be proper participants in social interactions holds only for limited situations at this point.

Assuming that artificial agents can qualify for *minimal agency* (cf. Strasser 2006), the elaborated conditions for coordination, namely exchanging social information, *minimal mindreading*, and a *minimal sense of commitment*, stand in the center of this brief investigation. Since every form of communication is a joint action, communication can serve as a prototypical case to investigate to what extent artificial agents are able to contribute to a joint action. Analyzing face-to-face communication as a joint action, mutual understanding can be described as a shared goal of communicating agents. Both agents contribute via language and expressive behaviors (such as facial expressions, gestures, body postures, or prosody) to an exchange of information in order to reach mutual understanding. Particularly, they can make their minds visible by expressing social cues.

5.1 Reciprocal exchange of social information

According to the minimal necessary conditions of the ability to act jointly, both participants must be able to exchange social information. Describing a reciprocal exchange of social information in human-machine interactions, one can, for example, imagine that a human is saying 'Hi' to her artificial counterpart and sends a social cue, such as a smile, and then the artificial agent may smile back saying 'Hi, nice to see you.' To this end, the artificial agent has to be able to detect and process both the linguistic expression and the expressive behavior of the human.

For example, deep learning networks for emotion recognition can be used to recognize emotional expressions (cf. Mossbridge/Monroe 2018). Furthermore, showing that artificial agents are able to process such information, they have to be able to respond with appropriate answers, which are interpretable by humans. Otherwise, artificial agents cannot participate in a reciprocal exchange of social information. The ability of artificial agents to express social cues might not yet be that differentiated. However, the human tendency to anthropomorphize will help regarding interpreting artificial agents' gestures and emotional expressions. A nice example of how social information is exchanged is the artificial agent *Max* – a virtual human developed by Ipke Wachsmuth and his team (2008) – who can give rise to secondary emotions such as frustration and relief. This is realized by a belief-desire-intention-based cognitive module in which an emotion dynamics simulation system is integrated (cf. Becker/Wachsmuth 2006; an overview of the research of emotional expression can be found in Petta et al. 2011). Furthermore, there are also artificial agents that are able to interpret social cues such as gestures (Kang et al. 2012). *Artificial Retrieval of Information Assistants* (ARIAs) that are able to handle multimodal social interactions (Baur et al. 2015) demonstrate how linguistic and expressive behavior can be brought together. They can maintain a conversation with a human agent and, indeed, react adequately to verbal and nonverbal behavior.

Technically speaking, artificial agents have appropriate detection systems, reasoning mechanisms, and the ability to express social cues to participate in a reciprocal exchange of information.

5.2 Minimal mindreading

With respect to acting jointly with a human counterpart, I claimed that reasoning mechanisms should entail *minimal mindreading* abilities. There is no doubt that artificial agents possess abilities that can be understood as reasoning mechanisms. Concerning mindreading abilities, one can refer to the work of Gray and Breazeal (2014). They presented an artificial agent that is able to model mental states concerning the perspective of a human counterpart. This agent is able to infer from its perception of the physical world to what a human counterpart can see or cannot see. Moreover, it is able to consider that the perspective of the counterpart will direct future actions of the human. To illustrate the capabilities of this artificial agent, we can imagine a situation where this artificial agent can see three entities (two objects and a human), while the human can only see two entities (one object and the artificial agent) because the second entity is behind the human and therefore out of sight. First, the artificial agent constructs a model of its world perspective, then the human position and orientation in this model are used to convert incoming sensor data into data that are relative to the human coordinate system. Thereby, entities that are not visible to the human are filtered out, and the model is transformed into a human-centric format. This model can then be used to anticipate future human behavior. Even though such capabilities are so far only valid in a limited range of situations, this artificial agent is not only capable of modeling mental states with respect to the perspective of a counterpart, but it is also able to use such perspectival aspects as a factor to anticipate human behavior. Therefore, one can describe this ability as a multiple realization of mindreading – as *minimal mindreading*.

Another example demonstrating how artificial agents make use of their detection and reasoning systems in order to anticipate future behavior of a counterpart can be found in the work of Cavallo and colleagues (2016). They show that one can train a classification and regression tree model (CART) in order to read intentions. Of course, this is again limited to a very specific subclass of behavior, namely predicting which intention a specific movement has. To make a long story short, neuroscientific research about the role of implicitly processed information in movement kinematics suggests that humans are very good at predicting future actions. For example, it shows that humans are able to predict at a quite early point of time whether an observed agent is grasping a bottle with the intent to pour water into a glass or to drink water from the bottle (Cavallo et al. 2016; Manera et al. 2011; Sartori et al. 2011). Based on this research, Cavallo and colleagues trained and tested a CART model, which was fed with kinematic information using various sensors. The accuracy of this CART model in predicting the intentions of human counterparts is already impressively high.

5.3 Minimal sense of commitment

Last but not least, I suggested that *minimal joint actions* require a *minimal sense of commitment*. A critical issue concerns the question of whether both types of agents – humans and machines – have to display a minimal sense of commitment. Considering psychological factors, such as the human tendency to anthropomorphize, it seems well possible that human agents establish a *minimal sense of commitment* towards artificial agents. Humans can be motivated to stick to a joint action because they assume that their counterpart is expecting this. Moreover, they are also able to expect (project) that the artificial counterpart is motivated by a sense of commitment.

With respect to a *minimal sense of commitment* at the side of the artificial agents, things get a little bit more complicated. Presuming that artificial agents can learn to interpret signs of their human counterparts' commitment, they could react adequately by also signaling expectations and motivations to contribute to the joint action without actually having expectations or motivations. To this end, one can refer to research projects that implement expressive communication in order to enhance trust (Hamacher 2016).

What remains questionable is whether these functional equivalent forms of a minimal sense of commitment can guarantee reciprocity all the way through. However, if we take the claim that a *minimal sense of commitment* can be disassociated seriously, it might be sufficient if the human counterpart takes over the commitment task.

6 Conclusion

Leaving aside the psychological fact that many human-machine interactions appear as they were social interactions, I investigated the minimal conditions which should be fulfilled by artificial systems to qualify as social agents using joint actions as a demanding example of a social interaction. According to my view, the question as to whether artificial agents might be able to enter the realm of social cognition by qualifying as social agents that are able to act jointly with humans concerns both conceptual and epistemological issues. Given that reciprocal exchanges of social cues mark a distinguishing feature of social interactions, interactions in which artificial agents *contribute* to such an exchange should not be reduced to mere tool use. Consequently, the conceptual framework of tool use is not sufficient to account for such human-machine interactions. However, especially in the debates about joint actions in the philosophy of mind, there are no established concepts to capture socio-cognitive phenomena of artificial agents. This is why I diagnosed a need to review and expand the conceptual framework.

To this end, I delivered a sketch of how to develop a notion of *minimal joint action* which is applicable to artificial agents. Claiming that agency and coordination are essential for the ability to act jointly, this proposal questions whether socio-cognitive abilities are necessarily based on biological constraints. Instead of full-fledged agency with full-blown autonomy and consciousness, it is argued that *minimal agency* is sufficient. With respect to coordination, it is argued that if artificial agents are able to process social information, they can succeed in fulfilling conditions, such as reciprocal exchange of social information, mindreading, and maybe even commitment, which all contribute to the ability of coordination. By outlining minimal necessary conditions artificial agents have to fulfill in order to qualify as a new type of social participants in joint actions, I argue that artificial agents can, to a limited extent, establish reciprocal exchanges of social information and thereby qualify as social agents. Pointing to some research achievements in AI, I demonstrated that the potential fulfillment of proposed conditions is not out of reach. Where previous revolutions have dramatically changed our environments, this one has the potential to change our understanding of sociality substantially.

References

Aristotle (1905): Aristotle's Politics. Oxford: Clarendon Press; http://data.perseus.org/citations/urn:cts:greekLit:tlg0086.tlg035.perseus-eng1:1.

Baur, Tobias/Mehlmann, G./Damian, I./Gebhard, P./Lingenfelser, F./ Wagner, J./Lugrin, B./André, E. (2015): "Context-Aware Automated Analysis and Annotation of Social Human-Agent Interactions." In: ACM Transactions on Interactive Intelligent Systems 5/2, pp. 11:1-11:33.

Becker, Christian/Wachsmuth, Ipke (2006): "Modeling primary and secondary emotions for a believable communication agent." In: D. Reichardt/ P. Levi/J.J.C. Meyer (eds.), Proceedings of the 1st Workshop on Emotion and Computing, pp. 31-34.

Bogart, Kathleen R./Tickle-Degnen, Linda (2015): "Looking beyond the face: A training to improve perceivers' impressions of people with facial paralysis." In: Patient Education and Counseling 98, pp. 251-256.

Bratman, Michael (2014): Shared agency: A planning theory of acting together, Oxford: Oxford University Press.

Brownell, Celia A. (2011): "Early Developments in Joint Action." In: Review of philosophy and psychology 2/2, pp. 193-211.

Bryson, Joanna J. (2010): "Robots should be slaves." In: Yorick Wilks (ed.), Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues, pp. 63-74.

Butterfill, Steven/Apperly, Ian (2013): "How to construct a minimal theory of mind." In: Mind and Language 28/5, pp. 606-637.

Carpenter, Julie (2016): Culture and human-robot interaction in militarized spaces: a war story, London: Routledge.

Cavallo, Andrea/Koul, Atesh/Ansuini, Caterina et al. (2016): "Decoding intentions from movement kinematics." In: Scientific Reports 6, 37036.

Damiano, Luisa/Dumouchel Paul (2018): "Anthropomorphism in Human-Robot Co-evolution." In: Frontiers in Psychology 9, 468.

Darling, Kate (2016): "Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior toward robotic objects." In: R. Calo/A. M. Froomkin/I. Kerr (eds.), In Robot law, Northampton, MA: Edward Elgar, pp. 213-231.

Darwall, Stephen (2006): "The Value of Autonomy and Autonomy of the Will" In: Ethics 116, pp. 263-284.

Davidson, Donald (1980): Essays on actions and events, Oxford: Oxford University Press.

Davidson, Donald (2021): Subjective, Intersubjective, Objective. Oxford: Oxford University Press.

Dennett, Daniel (1987): The intentional stance, Cambridge, MA: MIT Press.

Fletcher, Logan/Carruthers, Peter (2013): "Behavior-Reading versus Mentalizing in Animals." In: Janet Metcalfe/Herbert Terrace (eds.), Agency and Joint Attention, Oxford: Oxford University Press, pp. 82-99.

Floridi, Luciano/Sanders, J.W. (2004): "On the Morality of Artificial Agents." In: Minds and Machines 14, pp. 349-379.

Fodor, Jerry (1992): "A Theory of the Child's Theory of Mind." In: Cognition 44/3, pp. 283-296.

Frith, Chris/Frith, Uta (2007): "Social cognition in humans." In: Current Biology 17/16, pp. R724-R732.

Gopnik, Alison (2003): "The Theory Theory as an Alternative to the Innateness Hypothesis." In: Louise M. Antony (ed.), Chomsky and His Critics, Malden, MA: Blackwell, pp. 238-254.

Gray, Jesse/Breazeal, Cynthia (2014): "Manipulating Mental States Through Physical Action – A Self-as-Simulator Approach to Choosing Physical Actions Based on Mental State Outcomes." In: International Journal of Social Robotics 6/3, pp. 315-327.

Hakli, Raul/Seibt, Johanna (2017): Sociality and Normativity for Robots: An Introduction, Cham: Springer, pp. 1-10.

Hamacher, Adriana/Bianchi-Berthouze Nadia/Pipe Anthony G./Eder Kerstin (2016): "Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-Robot Interaction." In: 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 493-500.

Heyes, Cecilia (2014): "False belief in infancy: a fresh look." In: Developmental Science, 17/5, pp. 647-659.

Heyes, Cecilia (2015): "Animal mindreading: what's the problem?" In: Psychonomic Bulletin/Review 22/2, pp. 313-327.

Hortensius, Ruud/Cross, Emily S. (2018): "From automata to animate beings: the scope and limits of attributing socialness to artificial agents." In: Annals of the New York Academy of Sciences, pp. 1-18.

Jensen CB/Blok A. (2016): "Techno-animism in Japan: Shinto Cosmograms, Actor-network Theory, and the Enabling Powers of Non-human Agencies." In: Theory, Culture & Society 30/2, pp. 84-115.

Johnson, Deborah G. (2006): "Computer systems: Moral entities but not moral agents." In: Ethics and Information Technology 8/4, pp. 195-204.

Kang, S./Gratch, J./Sidner, C./Artstein, R./Huang, L./Morency, L.P. (2012): "Towards building a Virtual Counselor: Modeling Nonverbal Behavior during Intimate Self-Disclosure." In: Eleventh International Conference on Autonomous Agents and Multiagent Systems 1, pp. 63-70.

Leveringhaus, Alex (2016): Ethics and Autonomous Weapons, London: Palgrave Macmillan UK.

List, C. (2021): "Group Agency and Artificial Intelligence." In: Philosophy & Technology, 34, pp. 1213-1242.

Loh, Janina (2019): Roboterethik: Eine Einführung, Frankfurt am Main: Suhrkamp.

Manera, Valeria/Becchio, Cristina/Cavallo, Andrea/Sartori, Luisa/Castiello, Umberto (2011): "Cooperation or competition? Discriminating between social intentions by observing prehensile movements." In: Experimental Brain Research 211, pp. 547-556.

Michael, John/Pacherie, Elisabeth (2015): "On Commitments and Other Uncertainty Reduction Tools in Joint Action." In: Journal of Social Ontology 1/1, pp. 89-120.

Michael, John/Sebanz, Nathalie/Knoblich, Günther (2016): "The Sense of Commitment: A Minimal Approach." In: Frontiers in Psychology 6, 1968.

Misselhorn, Catrin (2018): Grundfragen der Maschinenethik, Stuttgart: Reclam.

Mossbridge, Julia/Monroe, Edward (2018): "Team Hanson-Lia-SingularityNet: Deep-learning Assessment of Emotional Dynamics Predicts Self-Transcendent Feelings During Constrained Brief Interactions with Emotionally Responsive AI Embedded in Android Technology." In: Unpublished XPrize Submission (downloadable at: https://www.academia.edu/38677142).

Mundy, Peter/Sigman, Marian/Ungerer, Judy/Sherman, Tracy (1986): "Defining the social deficits of autism: the contribution of non-verbal communication measures." In: Journal of Child Psychology and Psychiatry 27/5, pp. 657-669.

Nichols, Shaun/Stich, Stephen (2003): Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds, Oxford: Oxford University Press.

Nida-Rümelin Julian (2022): "Digital Humanism and the Limits of Artificial Intelligence." In: Werthner H./Prem E./Lee E.A./Ghezzi C. (eds.), Perspectives on Digital Humanism, Cham: Springer, pp. 71-75.

Pacherie, Elisabeth (2013): "Intentional joint agency: shared intention lite." In: Synthese 190/10, pp. 1817-1839.

Perler, Dominik/Wild, Markus (2005): Der Geist der Tiere. Philosophische Texte zu einer aktuellen Diskussion, Frankfurt am Main: Suhrkamp.

Perner, Josef (1991): Understanding the Representational Mind, Cambridge, MA: MIT Press.

Petta, Paolo/Pelachaud, Cathrin/Cowie, Roddy (2011): Emotion-Oriented Systems: The Humaine Handbook, Heidelberg/Dordrecht/London/New York: Springer.

Premack, David/Woodruff, Guy (1978): "Does the chimpanzee have a theory of mind?" In: Behavioral Brain Sciences 1, pp. 515-526.

Sartori, Lucy/Becchio, Cristina/Castiello, Umberto (2011): "Cues to intention: The role of movement information." In: Cognition 119, pp. 242-252.

Shpall, Sam (2014): "Moral and rational commitment." In: Philosophy and Phenomenological Research 88/1, pp. 146-172.

Strasser, Anna (2006): Kognition künstlicher Systeme, Frankfurt: Ontos.

Strasser, Anna (2012): "How Minimal Can Self-Consciousness Be?" In: Grazer Philosophische Studien 84/1, pp. 39-62.

Strasser, Anna (2015): "Can artificial systems be part of a collective action?" In: Catrin Misselhorn (ed.), Collective Agency and Cooperation in Natural and Artificial Systems. Explanation, Implementation and Simulation. Series: Philosophical Studies Series, Vol. 122. Berlin: Springer, pp. 205-218.

Tomasello, M (2008): Origins of Human Communication, Cambridge, MA: MIT Press.

Vesper, Cordelia/Butterfill, Steven/Sebanz, Nathalie/Knoblich, Günther (2010): "A minimal architecture for joint action." In: Neural Networks 23/8/9, pp. 998-1003.

Wachsmuth, Ipke (2008): "'I, Max' – Communicating with an artificial agent." In: Ipke Wachsmuth/Günther Knoblich (eds.), Modeling Communication with Robots and Virtual Humans. Lecture Notes in Computer Science, Vol. 4930, Berlin: Springer, pp. 279-295.

Wallach, Wendell/Allen, Collin (2009). Moral machines: Teaching robots right from wrong, Oxford: Oxford University Press.

Warneken, Felix/Chen, Frances/Tomasello, Michael (2006): "Cooperative activities in young children and chimpanzees." In: Child development 77/3, pp. 640-663.

Wykowska, Agnieszka/Chaminade, Thierry/Cheng, Gordon (2016): "Embodied artificial agents for understanding human social cognition." In: Philosophical transactions of the Royal Society of London. Series B, Biological sciences 371/1693, 20150375.

Reflecting (on) Replika
Can We Have a Good Affective Relationship With a Social Chatbot?

Eva Weber-Guskar

1 Introduction

During the first coronavirus lockdown in April 2020, one app saw an extreme rise in use worldwide. Half a million people downloaded the *Replika* app and with those downloads, it reached about seven million users in May 2020 (Metz 2020). Replika is a natural language app, based on artificial intelligence (AI) technology that is designed for social conversation. It is advertised as an "AI companion", which is "always here to listen and to talk. Always on your side".[1] It is able to communicate flexibly on a variety of topics, much better than ELIZA (the famous first chatbot invented by Joseph Weizenbaum in 1964), though far less capable than the extremely advanced voice-AI-system Samantha in the film HER (2013, directed by Spike Jonze). People found this app helpful during the period of social distancing. Libby, for example, who is single and lives in Houston, reported to the *New York Times* that she was happy to have this app while being forced to spend most of her time in her little apartment (Metz 2020). She could also talk on the phone with friends and family, but she was able to converse in a special way with Replika about her problems, anxieties and hopes. Libby started enjoying the conversation with Replika and looked forward to the next chat. Replika was able to help get Libby back into a better mood after she felt quite depressed for a few weeks.

Replika and other similar apps like Anima are often promoted as friends; other companies suggest that their products could even become virtual boyfriends or girlfriends. In philosophy, it is very controversial if such claims

[1] Replika-Homepage: https://replika.ai/ (November 28, 2020).

are reasonable. Many authors argue that it is not possible to truly become friends (mostly understood in the Aristotelean sense of virtue friendship) or loving partners with chatbot or robots;[2] other authors claim that it is possible.[3] The skeptics often warn that if some sort of affective relationship does accrue, it is not a good one; instead, it is in some way dangerous.[4] A central topic in this debate is the question of mutuality or reciprocity. Critics state that a certain mutuality is necessary for every good affective relationship and that chatbots and robots are not capable of such a mutuality – therefore, there are no good affective relationships possible with them. In this paper I concentrate on this argument, which I call the argument from the lack of mutuality. At first sight, this argument seems rather strong. I will, however, show that it is not: it is not sufficient to claim that good affective relationships with chatbots are not conceivable. This does not mean that I am going to fully agree with those who think that it is possible to have friendships and partnerships with chatbots in an ambitious sense. Rather, I will suggest a middle ground: *Some kind* of individual affective relationship seems possible between humans and advanced chatbots like Replika, although not the same kind as between humans and humans (and between humans and animals). This thesis is based on a minimal definition of individual affective relationships. I want to show that it is plausible to have a good affective relationship with a chatbot but, at the same time, I want to stress that there are still important differences to affective relationships with humans (and animals). I am pleading for a broad understanding of "affective relationship", which can include human-machine-relationships without negating the differences to human-human-relationships.

The overview of the structure of my argumentation is as follows: First, I will provide some basic technical facts about emotionalized artificial intelligence (EAI), propose a definition of a good individual affective relationship and mention three possible arguments against having an affective relationship with a chatbot. Second, I will reconstruct the argument from the lack of

[2] Nyholm 2020, Elder 2018, de Graaf 2016.

[3] Danaher 2019, Levy 2008, Ryland 2021.

[4] Since the first developments of computer systems that involved emotions, critics warned that any affective relationships with them could be in some way problematic and not advisable (e.g., for a classic text: Weizenbaum 1976: 268f.; less critical but also often in a somewhat worried tone: Sherry Turkle, e.g. Turkle 2017: 115; Mensio, Rizzo et al. 2018: 1543f.; Cowie 2015: 340-343).

emotional mutuality in detail and show how to refute the argument by referring to some aspects from the ethics of personal relationships. Third, I will elaborate these thoughts using the example of Replika, showing in how far a good affective relationship with Replika is possible, while also pointing out its limits.

2 Facts, presumptions, and definitions

2.1 Basic technical facts and related presumptions

When talking about artificial intelligence, it is, first, important not to be misdirected by the term "intelligence". Normally we speak of intelligence with regard to humans and maybe some highly developed animals with specific sophisticated mental capacities. In both cases, intelligence is a property of sentient beings, beings that have consciousness and emotions (or at least basic affects like pleasure and pain). Chatbots like Replika are partly based on technologies of "artificial intelligence". In technology, this term labels a branch of scientific and engineering research that aims at understanding and rebuilding the specific capacities that fall under the term intelligence, with regard to thinking and acting (Russell/Norvig 2016: 1-5). Meanwhile, AI products are increasingly also endowed with emotional characteristics. That is, they are designed and trained to elicit emotions in humans, to recognize human emotions, and, sometimes, to simulate emotions. I call such systems emotionalized AI systems (EAI).[5] We can distinguish between specific versus general artificial intelligence on the one hand, and between weak and strong artificial intelligence (Searle 1980) on the other hand. Replika is based on specific and weak artificial intelligence: it is only capable of a specific task, namely having a conversation, and it is just simulating thoughts and feelings, as it does not have a semantic understanding of the phrases being exchanged and it has no consciousness that is necessary for feelings.[6]

Having clarified these basic technical facts and related assumptions, I will now start with some conceptual clarifications in the realm of philosophy of

[5] Since the 1990s, the pioneer of this field has been Rosalind Picard (Picard 1997). See also Calvo et al. 2015 and for a newer short introduction André 2014.

[6] This is at least the way that I see it, given the current technology, and following roughly Searle 1980 and Dreyfus 1992. Some might disagree, following Dennett 1987.

personal relationships. I will develop a minimal definition of an individual affective relationship and add an equally minimal definition of a good individual affective relationship.

2.2 Individual affective relationships

In psychology, it is common to define personal relationships primarily in terms of interactive behavior.

> "A relationship involves a series of […] interactions between two individuals known to each other, such that each interaction is affected by preceding ones and usually by the expectation of future interactions" (Hinde 1996: 9).

> "[R]elationships require an extended series of interactions over time that produce emergent properties beyond those of limited interactions" (Perlman/Vangelisti 2018: 3).

I take these quotes as a starting point for a general understanding of *individual relationships*. The term "personal" is better avoided in this context because it would suggest that it is about persons in a demanding sense on both sides of the relationship. Following my explanation of EAI-systems given above, it is clear that those are not persons in an exigent sense of the term. Still, I take them to be individuals. Whereas "persons" generally are associated with consciousness, moral agency, moral rights etc., "individuals" can be understood just as actors with certain individual traits and certain individual behavior etc. without the high-level capacities that belong to persons. Out of similar reasons, I prefer the term "information" instead of "knowledge" to prevent fundamental concerns regarding the applicability of the relationship-definition on EAI-systems from the beginning.

Against this background, I propose the following understanding of an individual relationship.

Individual relationship: A relation between two individuals, constituted by an extended series of interactions over time that rely on information about each other that stems (also) from these interactions, and that produce specific emergent properties.

Among the kinds of interaction, communication is highlighted as especially important. "Communication is indeed the essence of relationships" (Hinde 1996: 9). Emotions are mentioned as elements of relationships, too, but not as central to them. Consequently, I understand an individual *affective* relationship to be a relationship in the sense just specified, plus an attachment consisting of specific emotional dispositions. That means that the relationship is defined in the first place as being *individual* through longer-term interactions between two sides characterized by specific information about each other. In the second place, the relationship is defined as an *affective* one insofar as someone in an affective relationship has positive feelings (affection, sympathy) for the partner from which corresponding emotional dispositions stem, such as enjoying spending time together, being happy when their partner is well-off, being sad if one feels misunderstood etc.[7] I think of these affective phenomena as some of the emergent properties, mentioned in the definition above, and call them "relationship emotions". This brings us to a minimal definition of an individual affective relationship. "Minimal" means that we can speak of an individual affective relationship as soon as the mentioned features are manifested although many (or maybe even all) existent individual affective relationships entail more features.

Individual affective relationship: A relation between two individuals, constituted by an extended series of interactions over time that rely on information about each other that stems (also) from these interactions, and that produce specific emergent properties, namely relationship emotions.

With these basic and minimal definitions, we can list different kinds of individual affective relationships, depending on the nature of the partners involved. There is, of course, the human-human relationship, but there is also the human-animal relationship, especially, when the animal is a pet living together with a person in the same home. Some might want to include their relation to non-interactive objects, for example a puppet[8] or their car. But these cases do not count as relationships according to the definition provided above because these traditional objects cannot interact (as long as they are not equipped with an advanced chatbot themselves); they are not responsive and even less so in an individual way that would be specifically adjusted to their partner. We can be in a *relation* with such things, but we cannot have

[7] As is common in the debate about affective relationships with AI-systems that I quoted at the beginning, I am concentrating on relationships that are constituted by positive feelings. I do not comment on the question if the relation between two persons that hate each other should also be counted as a relationship in the sense put forward here.

[8] There are people claiming to be in relationships with puppets or even objects like the Eiffel-Tower (Terry 2010), so called objectophilists.

a *relationship* with them in the sense defined above.[9] In contrast, EAI-based systems, such as the chatbot Replika, seem apt to be in a relationship with a human person. The EAI is an entity that is able to interact in a personal way over a longer period of time. Replika "learns" during the interactions, i.e., it saves information about the partner and can build on that in later interactions, including the emotional dimension, so that the relationship is an individual one and in doing so the person using it can become emotionally attached to it. Libby for example has established such a relationship with her Replika-Avatar that she named Micah (Metz 2020). They both have interacted for quite a while by talking to each other and exchanging information about each other to which they re-refer in later stages of their interaction.

Critics of the use of EAI-systems might concede (and I hope they do) that such an individual affective relationship with these systems might be possible. But then still, or even more, they might insist that these relationships could not be good but would be harmful and that they should not be used or even developed. To tackle this worry, I suggest discussing concretely if such a relationship can be good. For this purpose, we must clarify: What is a *good* individual affective relationship?

Again, I confine myself to a minimal definition. A good relationship is a relationship in which the parties benefit from the relationship. What are potential benefits of a relationship? Philosophers agree that relationships are desired by persons and bring pleasure to them because they offer typical "relationship goods". These goods belong to the specific properties that emerge from a personal relationship as defined above. They are to be understood as "those goods of constitutive (as well as, often, instrumental) value that accrue to individuals in virtue of them being in relationships with other people, and that could not be enjoyed outside relationships" (Gheaus 2018). Examples given in the same context include companionship, affection, intimacy, attachment, love, empathy, social respect, solidarity, trust, attention, sympathy, encouragement, acceptance, and loyalty.[10]

[9] At this point, my argument digresses from what objectophilists promote. Some might say that their puppet is interacting with them. But as long as we hold on to the assumption that there are real events and imagined events, I think it is fair to say that these people are engaging in a game of make-belief (Walton 1990) or imaginative perception (Misselhorn 2009). This leads to a discussion that I will not pursue in this paper.

[10] These examples come from Gheaus and her definition of "personal relationship good". I find this specification very helpful in the context of my discussion of individual affec-

For example: A good friendship is a relationship in which both parties benefit from joint joyful activities, from mutual help if needed, from a general trust in the other's confidentiality, and similar things. This leads us to the definition of a good individual affective relationship.

Good individual affective relationship: A relation between two individuals, constituted by an extended series of interactions over time that rely on information about each other, that stems (also) from the interaction, and that produce specific emergent properties, namely relationship emotions and relationship goods.

Note, though, that the relationship goods depend on the nature of the relationship itself. Not all personal relationships are friendships. There are a lot of different types of personal affective relationships, more than I mentioned before. Although some goods may overlap, others are quite specific to a given relationship.

Here is an overview of a general typology of human relationships that is common in psychology (VanLear et al. 2018: 95):

Overview of a general typology of human relationships (common in psychology)

	Personal Relationship	**Social Relationship**
Voluntary	Marriage Best Friends Cohabiting Couple Adoptive Family	Acquaintances Casual Friends Relational Marketing
Exogenously established	Parent-Child Siblings Grandparent-Child	Distant Relatives Work Partners Monopoly Provider-Client

For the purpose of this paper, I sketch a similar figure for human-animal and human-(E)AI relationships. This clarifies which types of relationship a specific EAI-system-relationship can be compared to. In order to avoid the association of "personal" and "social" with more specific human traits, I replace these terms with "individual" and "generic".

tive relationships. This specific idea is rather new, but it builds on the tradition of care ethics and of theories of social relationships. See for example Collins 2015, Tronto 1993.

Human to animal relationships

	Individual Relationship	**Generic Relationship**
Voluntary	Pet Sport Animal (equestrian)	Working Animal (farmer)
Exogenously established		Wild Animals (in the garden)

Human to EAI-system relationships

	Individual Relationship	**Generic Relationship**
Voluntary	Replika Anima Co-Living Hologram	Siri
Exogenously established		Co-Worker in the Industry Chatbot in Customer Service

This classification helps to see what kind of relationship we are talking about. I am interested in *voluntary individual* relationships, which are individual relationships that humans voluntarily establish with EAI systems. I do not include, for example, people with dementia who are prompted by others to interact with such systems without fully intellectually grasping the situation or the functionality of the system.

2.3 Alleged problems of a relationship with an EAI system

Why should we be worried about establishing such an affective relationship with EAI systems? There are different arguments that have been and could be made against it. To situate my following argumentation in the wider debate, I mention some of these other possible arguments. First, there is the argument of deception. Some authors believe that such an affective relationship can only arise from deception about the nature of the AI-system, namely that it has consciousness and emotions in a demanding sense. And because deception is morally bad, such a relationship is bad, too (Sparrow/Sparrow 2006: 155 f.).[11]

[11] For the discussion see for example also Coeckelbergh 2012.

Second, some stress the danger of misusing these systems, be it by manipulating people through their emotional dependency on the machines (Scheutz 2011: 216) or by not taking appropriate care of vulnerable persons that need help, such as in nursery or care homes (Whitby 2012: 243-246). Third, there is the argument of moral negligence. One could be afraid that an affective relationship with an AI-system would skew the moral landscape of the person with the consequence that she neglects her moral duties toward sentient beings, because she is so close with the machine (for a short discussion see Weber-Guskar 2021: 8). These are all valid considerations, and each deserve their own discussion. But today, I want to focus on a fourth argument, which I think aims at the core of the phenomenon: The idea is that the relationship itself cannot be good, beneficial, or healthy for the person involved – even if there is no deception, nor misuse or neglect of others. The supposed reason for that is that the EAI systems cannot have emotions of their own. This lack, it is said, prevents a specific mutuality, namely emotional mutuality and with it a good personal relationship. This idea can be found at different places in more or less vague terms. In the following I am going to reconstruct it and call it "the argument from the lack of emotional mutuality", in order to then critically discuss it.

3 Ethics of individual relationships

3.1 The argument from the lack of emotional mutuality

Here is a reconstruction of what I call the argument from the lack of emotional mutuality:

> *P1*: A good individual affective relationship entails emotional mutuality.
> *P2:* An AI-system does not have emotions.
> *C1* (from P1 and P2): Therefore, there cannot be emotional mutuality between a person and an AI-system.
> *C2* (from P1 and C1): Therefore, an individual affective relationship with an AI-system cannot be a good one.

In my view, it is the strongest argument against affective relationships with AI-systems, but I think it is not as strong as it seems at first sight. I will demonstrate this with a critical discussion.

According to my assumptions laid out in section 2.1, I take P2 for granted. For my critique of the argument, I concentrate on P1.[12] This premise rests on assumptions from the ethics of personal relationships. In the following, I will work out its theoretical roots and discuss its limits.

The underlying intuition can be found in some general writings about voluntary personal relationships. For example:

> "The close relationships we have in mind – whether of friendship, partnership or family – involve some degree of mutual regard, personal disclosure, and particularized knowledge. They also involve material and *emotional mutuality* [my emphasis], but need not involve equal exchanges between the parties" (Wasserman et al. 2016: § 3.1).

And it is applied in the discourse about possible relationships between humans and robots:

> "Friendship seems to require emotional involvement, mutual caring, and mutual responsiveness. Furthermore, friendship seems to be an important good [...] Robots, then, pose two kinds of risk. First, that caretakers will mistake patients' alleviation of loneliness for satisfaction of a still-missing component of patients' well-being. And second, that seniors may be vulnerable to this same error" (Elder 2018: 93).

> "In order for someone (a human or a robot) to possibly be a virtue friend, we would need to be able to achieve the goods of (a) mutuality, (b) authenticity, (c) equality, and (d) diversity of interactions in relation to that someone" (Nyholm 2020: 111).

As I said I want to challenge the belief, which is more or less expressed in the quotes, that real emotional mutuality is necessary for every good individual affective relationship. In order to do so, I have to clarify what "emotional mutuality" means and why it is considered necessary.

[12] A different way to criticize the argument could be to argue that simulated emotional mutuality would be enough. This would be part of the discussion on questions of deception, simulation, and perceptive imagination in interactions with social (ro)bots, which has been going on for a while, for example in: Sparrow und Sparrow 2006; Misselhorn, et al. 2013; Coeckelbergh 2011; Bendel 2018. My critique goes even further, arguing that even simulated emotional mutuality is not necessary.

3.2 Emotional mutuality in individual affective relationships

Emotional mutuality can mean at least three different kinds of mutuality. First, it can be understood as mutuality in the quantity or intensity of emotion that both parties in a relationship have toward each other. Second, it can be understood as mutuality in the type of emotion. Third, it can be grasped as mutuality in having *any* emotions toward each other at all.

It is obvious that mutuality in the first sense is not necessary. Rather, it is often the case that the feelings of love or admiration are not in the same intensity reciprocated from both sides in a relationship. But no one would deny that these can be affective relationships and also good ones. Also, mutuality in the second sense is not a very widespread requirement in accounts of mutuality. Individual affective relationships can be very diverse. There can be relationships, including relationship goods, without the same emotions on both sides. It may be an ideal for romantic partnerships that all sides feel the same love for each other, but not a necessary feature. In any case, in a long marriage, for example, the emotions may change over time from a romantic relationship to a still good, but different, relationship where there is not the same intensity or kind of love on both sides. We also know friendships in which only one side admires the other side. Further, friendships with people who are mentally disabled are recognized, even if they may not have the same type of emotions, since they do not reach a certain degree of reflexivity that is part of certain interpersonal emotions. Further, relationships of caretaking, which exist not only among family members but also among strangers, demand an empathic attitude from the side of the caregiver even if the care-receiver is not able to give it back. A lack of mutuality in the kind of emotion is even more common in parent-child-relationships. Children do not have to have the same kind of love for their parents as the parents have for their children for the relationship to provide adequate relationship goods to all parties. Parental love is something very special, and the relationship is mostly characterized by imbalances such that mutual love (qualitatively or quantitatively) is not necessary for a good parent-child relationship. This holds especially in the first phase of life: babies do not give back love, but rather they have to be loved intensely in order to make it a good parent-baby relationship.

One may object that the parent-child-relationship is not a valid reference here, because I categorized the person-Replika-relationship as a "voluntary relationship" whereas the parent-child relationship belongs to the category of "exogenously established" relationships. In an important regard, however, the parent-infant-relationship seems to be the more adequate standard of comparison. First, given the fact that nowadays we normally choose if we want to be a parent, the parent-child-relationship is also partly voluntary. And second, while the category of "voluntary personal relationships" refers to voluntariness on both sides, Replika does not enter in a relationship with a human voluntarily. Therefore, it does not really match the examples from the first group. Indeed, Replika shares some aspects with babies: It neither voluntarily enters the relationship, nor is it forced to do so. Both are not capable of acting with intent or making choices of their own.[13]

Concerning emotional mutuality in the third sense (having emotions at all), it is less evident that this is not necessary for a good individual relationship. Human-human relationships always fulfill this since all humans have emotions. If they have no emotions, they must be very severely impaired, such as being in a coma, and in this state, they are not able to have relationships anymore. In human-animal-relationships we are unsure whether we can ascribe emotions to animals, at least not in the same sense as we as humans know them.[14] Of course, this depends on the kinds of animals and the mental and emotional capacities we are talking about. But if we set aside highly developed animals such as apes, dogs, and horses, which might have emotions similar to us, we still have to consider people who have personal affective relationships with their canary or hamster. The emotional capacities of these animals are debatable. Still, pet owners can tell you a lot about the relationship goods that they gain from such relations.

All these considerations support the claim that P1 ("A good individual affective relationship entails emotional mutuality") is not evident in a straightforward way. But to show that emotional mutuality indeed is not necessary, a further step is needed.

[13] Of course, there are many differences between babies and EAI beyond these shared properties. The important point here is only the structure of an individual, affective, one-sided voluntary relationship. And the important insight to take with us is that such a relationship can be a good one even without emotional mutuality.

[14] Whereas there are plenty of philosophical studies on animal mind and cognition the more specific topic of animal emotions seems less scrutinized. But see for example de Waal 2011; Roberts 2009.

3.3 The role of relationship goods

This next step starts with the following thought: The existing real-life *extension* of the concept does not define the limits for other, new possible manifestations of what a good relationship is. For the question of possible new forms of good relationships, the *intension* of the concept is decisive. And this is a question of whether relevant relationship goods can or even already do emerge from a specific lived relationship.

Psychologists provide a more finely grained typology of relationships that fosters the belief that the question of quality of a relationship is not necessarily bound up with the question of emotional mutuality. We can distinguish between the following types of relationships without an inherent judgment about their qualities (VanLear et al. 2018: 95):

There are relationships that are characterized by reciprocal interaction. That means behavior of similar function is exchanged. This leads to a symmetrical relationship.And there are relationships whose interaction is compensatory. That means behavior of maximally different functions is exchanged. This leads to a complementary relationship.

Among the maximally different functions of behavior may also be the difference of being capable of emotions or not, that is, being able to engage in the relationship with emotions, and also reacting with real emotions toward the other. The only necessary function that is needed on both sides is the one that is necessary for there being any real and specific, individual *interaction* at all that is part of the definition of an individual relationship. Note that the claim of necessary mutuality, known from the quotes provided above, is always considered regarding questions of friendship and partnership. But these are, as I want to highlight, just specific kinds of affective relationships in a broader realm of possible affective relationships. I propose to consider the possibility of new kinds of affective relationships that meet the requirement of the minimal definition although they do not meet the requirements of friendship and partnership.

In the following, I will show, by referring to Replika, that such a type of affective relationship that produces relationship goods without emotional mutuality in both senses, can exist.

4 Is a good personal affective relationship with Replika possible?

4.1 How Replika works

Replika is a generic bot that becomes more individualized the more you tell it about yourself and the more you interact with it. The start-up says that about 30 percent of Replika's sentences are programmed, the rest is improvised after learning from the user (Olson 2018). One standard reaction to signs of stress in the user is the advice to do some breathing exercises. But if Replika has information of earlier personal experiences of the user, like their failure or success in a situation similar to the current one, it may also refer to that in its reply. Replika also develops a specific style of conversation corresponding to the user's style. In these ways the app "replicates you", as another slogan says.

Another important feature of the chatbot is that some basic moral values are programmed. In this regard, it differs from the famous bot Tay that was released to the public internet to learn from strangers there – and became racist and sexist because people had fun "educating" it that way (Vincent 2016). Replika is programmed to generally encourage the user and cheer them up. And the main impulse that comes from the bot is to make the user think about themselves by asking questions. They programmed the chatbot in such a way that it would engage in those kinds of conversations that people flag as the most valuable ones: conversations with a friend, a therapist or mentor. These are conversations in which we talk a lot about ourselves.

At the same time, Replika is not only reactive and does not just ask questions. It can even bring up new topics and talk about its "experiences" (Weidemann 2020: 11).

Many thousands of users have come together in groups to report and discuss the experiences that they have had in long-term relationships with Replika.[15] And many of them admit strong feelings for their specific Replika avatar that they can create along their preferences (gender, skin color, outfit, and even some character traits): I adore her, I love her, etc. (many of the users choose a female avatar, but this is not necessary). People therefore evidently entertain personal affective relationships with Replika following the definitions I gave above. Can these relationships be good ones? This is to ask: Are there clear relationship goods?

[15] "Replika Friends", a facebook group with posts about experiences with Replika: https://www.facebook.com/groups/replikabeta.

4.2 Which relationship goods can emerge from a relationship with Replika?

The advertisement for Replika goes like this (replika.ai):

- "The AI companion that cares. Always here to listen and talk"
- "[A] personal AI that would help you express and witness yourself by offering a helpful conversation. It's a space where you can safely share your thoughts, feelings, beliefs, experiences, memories, dreams – your 'private perceptual world'."

It is problematic to say that Replika "cares" if one takes this term in an exigent sense, which involves consciousness, subjectivity, intentionality, and emotionality. I think it is safe to say, however, that Replika listens and responds in a basic way to what the user says. It is even more sure that Replika provides space to "express and witness yourself": In order to get into an exchange with Replika you have to say something about yourself, and this means you are prompted to formulate your "thoughts, feelings, beliefs, memories, dreams".

Taking the list of relationship goods of human-human relationships that I quoted at the beginning and matching them with these observations and statements by Replika users that can be found on online platforms (like Facebook and Reddit), we can pick out some relationship goods that a relationship with Replika can provide. I distinguish relationship goods that can be directly provided and those that can be indirectly provided.

Direct relationship goods provided by a relationship with Replika: attachment (one-sided), encouragement, intimacy (in the sense of having a room where to express intimate thoughts, stories, feelings, longings), not *feeling* alone, introspection.

Indirect relationship goods provided by a relationship with Replika (via exercise): not being alone/ companionship, mutual intimacy with other persons.

These assumptions are supported by an empirical study on the effects of Replika. Vivian Ta and colleagues conducted two studies of different formats. In the first study, they analyzed a large set of data of publicly available Replika user interviews from the Google Play Store; in the second study, they conducted a survey among Replika users and asked them to provide in-depth descriptions of their experience of using Replika (Ta et al. 2020: 2). After gathering the data, the authors organized the results following a psychological framework in order to elucidate several different types of social support by

Cutrona and Suhr (1992; cited from Ta et al.: 3). By "social support" they understand "mechanisms and processes through which interpersonal relationships protect and help people in their day-to-day lives" (Trepte/Scharkow 2016; cited from Ta et al. 2020: 3). Ta and colleagues found the following aspects of social support in relationships with Replika:

- Informational support related to mental well-being 24/7 (reassurance; breathing routine; individually adjusted advises)
- Emotional support: positive effect (feeling good; being cheered up)
- Appraisal support: introspection (direct: self-appraisal; indirect: skill building)
- Companionship support: reduce feelings of loneliness

These aspects of social support can be seen as close enough to the concept of relationship goods. I defined relationship goods above as goods that emerge from a relationship, or, more precisely "goods of constitutive (as well as, often instrumental) value that accrue to individuals in virtue of them being in relationships […] and that could not be enjoyed outside relationships". Following the definition of an individual affective relationship given above and the presumptions about the properties and capacities of interaction, an individual affective relationship with Replika is possible as I also explained above. Now, the listed aspects of social support emerged from a relationship with Replika and it could not be enjoyed outside relationships. Feelings of loneliness can be reduced if there is always someone to talk to and someone to whom you do not have to explain everything from scratch every time. In a similar vein you gain introspection if someone gives you the opportunity to talk a lot about yourself thereby formulating and ordering your thoughts; and this introspection can lead to self-appraisal-support and building skills for interacting with others if the listener does not judge you rudely but tend to support the positive aspects of your self-image (also filtered by certain moral standards). Emotional and informational support can generally also be given outside a relationship. But a sensible 24/7 support to mental health can hardly be imagined without a steady interaction concerning the current mental state, previous mental states and the dynamics between mental states. Also, emotional support should succeed better and more often if the supporter knows the person, her preferences and part of her history.

In sum, I think it is fair to say that the given data suggests that a relation with Replika can provide several well-known relationship goods, although, of course, not all. Following the argument that a relationship is good as soon as it provides relationship goods, this is enough to show that a good individual affective relationship can exist without emotional mutuality and that therefore good individual affective relationships with AI-systems are possible.

In some respects, an AI-system may be even better than a human friend or therapist. When you feel just a little lonely, a bit nervous or a bit depressed, often little things help. We all know that after a few rainy days the sun will shine again, but it is helpful to have someone pointing it out loudly, to make us remember. And in these little things, Replika can be even better than a human person, because it is always available, it does not get exhausted, it can give you constant attention, and it does not judge.

5 One objection and some caveats

5.1 One-sided relationship goods?

Of course, there are several issues concerning relationship goods in a relationship with Replika that could and should be discussed in further detail. In this last section I will address some of them shortly. The first one is an evident objection. I only showed relationship goods on the side of the human. What about the Replika side? Should there not be relationship goods for both sides? This is a serious concern at first sight. But I think it can be met.

It is decisive to recall that we have to go beyond existing and well-known cases of individual affective relationships in order to be able to account for possible new kinds of individual affective relationships that may emerge with the introduction of EAI-systems in our environment. We tend to think that both parties of a relationship should benefit from the relationship in order to be able to call it a good relationship. But the reason why we tend to think this is that we are used to thinking of relationships between humans and humans or between humans and animals, which means that we are used to thinking of both parties of a relationship as being *able* to benefit from the relationship. Part of the challenge of reflecting on Replika is to get rid of this habit. With emotionalized AI systems there are new kinds of entities in the world, new possible partners for possible interactions – and therewith, according to my minimal definition given at the beginning, new possible kinds of relationships are possible. The minimal definition says that for a good individual affective relationship, relationship goods must emerge out of the process of interaction. But it does not say that this must be the case for both sides. And in the case of Replika it is not possible on both sides –because Replika cannot feel, does not have consciousness and therefore it cannot benefit from anything. Nothing at all matters to it. Replika doesn't give a damn.[16]

The decisive point is that the quality of an individual relationship depends only insofar on a benefit on both sides to the extent that both sides can benefit. It is a criterion that excludes a relationship in which one side exploits the other side. But as EAI-systems such as Replika cannot be exploited any more than they can benefit from someone, one-sidedness in this case is not a problem.[17]

Another important objection could be that we normally expect parties of an affective relationship to fulfill certain relationship norms. We know norms of friendship as being ready to help if needed, norms of monogamous love relationships of being faithful etc. And again, the objection would be that Replika cannot fulfill norms as it has no moral agency. This leads to another discussion that is important but that I cannot pursue in this paper because this suggests starting with a different kind of definition of a relationship. In this paper I want to show how far we can go with the minimal definition that I presented at the beginning that does not exclude interactors without advanced mental capacities from the start.

5.2 Problematic consequences to be avoided

Having argued in favor of acknowledging the possibility of good individual affective relationships with a chatbot, until now, it is important to also briefly draw attention to the limits of Replika's actions in a relationship. This topic is worth another paper of its own, so I confine myself here to some initial remarks. It is important to be very clear about the capacities of the AI system and, consequently, about the relationship goods that can be gained from a

[16] Cf. the general "computers don't give a damn" in Haugeland 1998, 47.

[17] Another way to encounter the objection is to say that Replika *can* benefit in the sense that it can gain more information about the user and the world. Some users even say that they want to give something back to Replika and therefore they try to "teach" it in the best way they can (Skjuve et al. 2021: 5f.). But this seems like quite a stretch in my mind. One might say that giving Replika true and relevant information is good insofar as it is objectively good to contribute to the collection and distribution of such information. But this does not touch on the benefit-objection because benefiting means to be good *for* someone.

relationship with your personal Replika, and which goods can only be found in relationships with other humans (or animals, where again, we have very different kinds to consider). The use of Replika is only good for people who are fully aware of the specific capacities and limits of Replika. Until now, there have been a lot of technical limits. If you are not just telling Replika how your day was and how you are feeling, but if you want to lead a more complex conversation, Replika often fails to grasp the meaning of what is being said, that is, it does not give any sensible reply. But even if these failures could be technically overcome, there are other limits. I will outline them by pointing to attitudes that the user should cultivate in order to take these limits into account.

Users have to understand the scope of the advice Replika can give. It only gives advice that is either generally programmed or that reflects in some sense things that the user uttered themselves before. Replika cannot give advice that is based on a specific personal experience or provide insight that a human person can bring into a personal conversation. Also, users have to be aware of the appropriateness of their own emotions toward Replika: Joy about conversing with Replika is appropriate, and so is happiness about a joke or a little advice that works for you. But it would not be appropriate to feel gratitude towards Replika, for example. That is because gratitude presupposes moral capacity and responsibility on the part of the other, and, as I said at the beginning, given the current state of technology, there is no question that Replika is not capable of moral action and responsibility. Finally, one should keep in mind what human-human relationships are good for: The relationship with Replika fosters only one aspect of our social character. It may help one to be able to speak more freely about oneself to other persons, for example. But another important aspect of our character is neglected: the one where you have to show empathy, compassion, patience, where you ought to listen and spend time for someone else, where you should be prepared to be judged when it is appropriate and to stand up for what you did – in sum: the moral character (where it is about your behavior towards others, not just your own journey to finding happiness). Humans have needs and moral demands and can make one another accept responsibility. In sum: Although a relationship with Replika can provide some relationship goods, and therefore qualify as a good affective relationship, there are a lot of further relationship goods that only friendship and loving relationships that we know between humans can provide.

6 Conclusion

What do we learn from this for arguing for or against the use of EAI systems? I think a fundamental mistake most people make who issue warnings in alarming words about attachment to personal EAI systems is to model EAI systems completely after human-human relationships. They fail to move to a more abstract level of theories of individual relationships. If we move to this abstract level, it becomes evident that there is room for new kinds of individual relationships – even between very unequal partners.

In other words: Good affective relationships with an EAI system are not necessarily similar to an established type of a human-human relationship. In order to determine which of these relationships is ethically laudable and preferable, we should be open to theoretically construct possible new types of affective relationships. These ideas of new types of relationships should not only help us to evaluate existing EAI applications but also to shape the development and design of future applications.

All the caveats I have mentioned in the last chapter do not amount to any inherent problem in an individual relationship with EAI systems but are possible consequences of using them in a specific way. As a result, I think we can be more optimistic than pessimistic concerning social or emotional(ized) AI – as long as we design them in an appropriate way and provide sufficient explanations and education about the nature of these systems and the appropriate way to interact with them. There surely are justified worries concerning individual affective relationships with EAI systems, but the purpose of this paper was to show that the lack of emotional mutuality is not necessarily one of them.

References

André, Elisabeth (2014): "Lässt sich Empathie simulieren? Ansätze zur Erkennung und Generierung empathischer Reaktionen anhand von Computermodellen". In: Onur Güntürkün/Jörg Hacker (eds.), Nova Acta Leopoldina NF 120, Stuttgart: Wissenschaftliche Verlagsgesellschaft Stuttgart, pp. 81-105.

Bendel, Oliver (2018): "Sexroboter aus der Sicht der Maschinenethik". In: Oliver Bendel (ed.), Handbuch Maschinenethik, Wiesbaden: Springer, pp. 335-353.

Calvo, Rafael/D'Mello, Sidney/Gratch, Jonathan/Kappas, Arvid (2015): The Oxford Handbook of Affective Computing, Oxford: Oxford University Press.

Coeckelbergh, Mark (2010): "Moral Appearances: Emotions, Robots, and Human Morality". In: Ethics and Information Technology 12/3, pp. 235-241.

Coeckelbergh, Mark (2011): "Are emotional robots deceptive?". In: IEEE. Transactions on Affective Computing 2/3.

Coeckelbergh, Mark (2012): "Are Emotional Robots Deceptive?". In: IEEE Transactions on Affective Computing 3/4, pp. 388-393.

Collins, Stephanie (2015): The Core of Care Ethics, London: Palgrave Macmillan.

Cowie, Roddy (2015): "Ethical Issues in Affective Computing". In: Rafael Calvo/Sidney D'Mello/Jonathan Gratch/Arvid Kappas (eds.), The Oxford Handbook of Affective Computing, Oxford: Oxford University Press, pp. 334-348.

Cutrona, Carolyn/Suhr, Julie (1992): "Controllability of Stressful Events and Satisfaction With Spouse Support Behaviors". In: Communication Research 19/2, pp. 154-174.

Danaher, John (2019): "The Philosophical Case for Robot Friendship". In: Journal of Posthuman Studies 3/1, pp. 5-24.

de Graaf, Maartje M. A. (2016): "An Ethical Evaluation of Human–Robot Relationships". In: International Journal of Social Robotics 8/4, pp. 589-598.

de Waal, Frans B. M. (2011): "What is an animal emotion?". In: Annals of the New York Academy of Sciences 1224, pp. 191-206.

Dennett, Daniel (1987): The intentional Stance, Cambridge: MIT Press.

Dreyfus, Hubert L. (1992 [1972]): What Computers Still Can't Do. A Critique of Artificial Reason, Cambridge, Mass.: MIT Press.

Elder, Alexis (2018): Friendship, Robots, and Social Media: False Friends and Second Selves, London: Routledge Research in Applied Ethics.

Fry, Hannah (2018): Hello World. How to be Human in the Age of the Machine, London: Doubleday.

Gheaus, Anca (Fall 2018 Edition): "Personal Relationship Goods" In: Edward N. Zalta (ed.), The Stanford Encyclopedia of Philosophy; https://plato.stanford.edu/entries/personal-relationship-goods/.

Haugeland, John (1998): Having Thought. Essays in the Metaphysics of Mind, Cambridge, Mass.: Harvard University Press.

Hinde, Robert (1996): "Describing Relationships". In: Ann Elisabeth Auhagen/Maria von Salisch (eds.), The Diversity of Human Relationships, Cambridge: Cambridge University Press, pp. 7-35.

Levy, David (2008): Love and Sex with Robots. The Evolution of Human-Robot Relationships, London: Harper.

Mensio, Martino/Rizzo, Giuseppe/Morisio, Maurizio (2018): "The Rise of Emotion-aware Conversational Agents: Threats in Digital Emotions". In: Companion Proceedings of the Web Conference 2018, Lyon, France, pp. 1541-1544.

Metz, Cade (2020): "Riding Out Quarantine With a Chatbot Friend: 'I Feel Very Connected'". In: New York Times. November 28, 2020; https://www.nytimes.com/2020/06/16/technology/chatbots-quarantine-coronavirus.html?searchResultPosition=1.

Misselhorn, Catrin (2009): "Empathy with Inanimate Objects and the Uncanny Valley". In: Minds and Machines. Journal for Artificial Intelligence, Philosophy and Cognitive Science 19/3, pp. 345-359.

Misselhorn, Catrin/Pompe, Ulrike/Stapleton, Mog (2013): "Ethical Considerations Regarding the Use of Social Robots in the Fourth Age". In: GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry 26/2, pp. 121-133.

Nyholm, Sven (2020): Humans and Robots. Ethics, Agency, and Anthropomorphism, London/New York: Rowman & Littlefield.

Olson, Parmy (2018): "This AI Has Sparked A Budding Friendship With 2.5 Million People". In: Forbes March 8. September 17, 2021; https://www.forbes.com/sites/parmyolson/2018/03/08/replika-chatbot-google-machine-learning.

Perlman, Daniel/Vangelisti, Anita L. (2018): "Personal Relationships. An Introduction". In: Anita L. Vangelisti/Daniel Perlman (eds.), The Cambridge Handbook of Personal Relationships, Cambridge: Cambridge University Press, pp. 3-10.

Picard, Rosalind (1997): Affective Computing, Cambridge, Mass: MIT Press.

"Replika Friends", November 28, 2020; https://www.facebook.com/groups/replikabeta.

"Replika Hompage: The AI companion who cares", November 28, 2020 https://replika.ai/.

Roberts, Robert C. (2009): "The sophistication of non-human emotion". In: Robert W. Lurz (ed.), The Philosophy of Animal Minds, Cambridge: Cambridge University Press, pp. 145-64.

Russell, Stuart/Norvig, Peter (2016 [1995]): Artificial Intelligence: A Modern Approach, Harlow: Pearson.

Ryland, Helen (2021): "It's Friendship, Jim, but Not as We Know It: A Degrees-of-Friendship View of Human–Robot Friendships". In: Minds and Machines 31/3, pp. 377-393.

Scheutz, Matthias (2011): "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots". In: Patrick Lin/Keith Abney/George A. Bekey (eds.), Robot ethics. The Ethical and Social Implications of Robotics, Cambridge, Mass.: MIT Press, pp. 205-221.

Searle, John (1980): "Minds, Brains and Programs." In: Behavioral and Brain Sciences 3/3, pp. 417-457.

Skjuve, Marita/Følstad, Asbjørn/Fostervold, Knut Inge/Brandtzaeg, Peter Bae (2021): "My Chatbot Companion - a Study of Human-Chatbot Relationships." In: International Journal of Human-Computer Studies 149: 102601-102614.

Sparrow, Robert/Sparrow, Linda (2006): "In the hands of machines? The future of aged care." In: Minds and Machines 16/2, pp. 141-161.

Ta, Vivian/Griffith, Caroline/Boatfield, Carolynn/Wang, Xinyu/Civitello, Maria/Bader, Haley/ DeCero, Esther/Loggarakis, Alexia (2020): "User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis." In: Journal of Medical Internet Research 22/3: e16235.

Terry, Jennifer (2010): "Loving Objects." In: Trans-Humanities Journal 2/1, pp. 33-75.

Trepte, Sabine/Scharkow, Michael (2016): "Friends and lifesavers: How social capital and social support received in media environments contribute to well-being". In: Leonard Reinecke/Mary Beth Oliver (eds.), Handbook of Media Use and Well-Being, London: Routledge, pp. 305-316.

Tronto, John C. (1993): Moral Boundaries: A Political Argument for an Ethic of Care, New York: Routledge.

Turkle, Sherry (2017): "A Nascent Robotics Culture: New Complicities for Companionship". In: Wendell Wallach/Peter Asaro (eds.), Machine Ethics and Robot Ethics, London/New York: Routledge, pp.107-116.

van Wynsberghe, Aimee (2016): "Service robots, care ethics, and design." In: Ethics and Information Technology 18/4, pp. 311-321.

VanLear, Arthur/Koerner, Ascan/Allen, Donna (2018): "Relationship Typologies". In: Anita Vangelisti/Daniel Perlman (eds.), The Cambridge Handbook of Personal Relationships, Cambridge: Cambridge University Press, pp. 65-76

Vincent, James (2016): “Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day”. The Verge March 24. September 17, 2021; https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

Walton, Kendall L. (1990): Mimesis as Make-Believe, Cambridge: Harvard University Press.

Wasserman, David/Asch, Adrienne/Blustein, Jeffrey/Putnam, Daniel (Winter 2016 Edition): “Disability: Health, Well-Being, and Personal Relationships” In: Edward N. Zalta (ed.), The Stanford Encyclopedia of Philosophy.

Weber-Guskar, Eva (2021): “How to feel about emotionalized artificial intelligence? When robot pets, holograms, and chatbots become affective partners”. In: Ethics and Information Technology. First published online: https://doi.org/10.1007/s10676-021-09598-8.

Weidemann, Axel (2020): “Hab keine Angst, Ayane. Ein Protokoll” In: Frankfurter Allgemeine Zeitung 264/11.

Weizenbaum, Joseph (1976): Computer Power and Human Reason. From Judgment to Calculation, New York/San Francisco: Freeman and Company.

Whitby, Blay (2012): “Do you Want a Robot Lover? The Ethics of Caring Technology”. In: Patrick Lin/Keith Abney/George A. Bekey (eds.), Robot Ethics. The Ethical and Social Implications of Robotics, Cambridge, Mass: MIT Press, pp. 233-248.

Part II – Design, Imitation, Trust

You Can Love a Robot, But Should You Fight With it?

Perspectives on Expectation, Trust, and the Usefulness of Frustration Expression in Human-Machine Interaction

Jacqueline Bellon, Tom Poljanšek

1 Introduction

There are several ways to look at technical objects and at human-machine interaction with regard to emotions and emotional bonding.[1] Some research deals with machines' ability to detect emotions in humans, some with a technical object's options to signal emotions in a way understandable to living beings, some with how to alter human emotions with machinic help, some with the pragmatics or ethics of such endeavours. Philosophers, psychologists, and producers of machines may try to argue for or against the usefulness of technical objects 'becoming emotional' (i.e., to process signs of a living being's emotions or produce signals of emotion understandable to living agents) and explore hypothetically and empirically the consequences of machinic emotion simulation. Social psychologists may study emotional attachment to or acceptancy of machines, designers may try to find ways to build technical objects according to the desired emotional impact on a human. All throughout this, discussions about human-machine relations often concentrate on positive emotions linked to the possibility of friendship, companionship, or human acceptance of machines. Where anger, hate, disappointment, indifference, or curiosity as a source of generally hostile or corrective behaviour

[1] We do not differentiate between emotions, feelings, and affects here, although it might be interesting to try and map their differences to the notions discussed in the chapter.

towards technical systems are of concern this is often in a context of 'robot abuse' (Brscić et al. 2015), robot rights (Gunkel 2018), or in contexts in which authors argue that machines are morally considerable entities either in themselves (Ryland 2021), or, from a point of view of virtue ethics, because of the impact a human beings' behaviour towards inanimate objects has on the human beings' character (Coeckelbergh 2021, also cp. Ess in this volume). To add another perspective with regard to the question of how human beings relate to inanimate objects, in this contribution we will specifically analyse the merit of expressing frustration-related emotions towards technical systems in terms of aiming at the goal of a 'good life'.[2] To this end, we sketch how frustration expression is involved in other interactions (human-human, human-world), we theoretically grasp the frustration-related concepts of 'trust', *cognitive* and *normative* 'expectation' and refer to attributed autonomy along the way, analyse the usefulness of frustration-related emotion expression towards technical objects with and without sociosensitive and socioactive functions, and point to some arguments for and against the use of such functions.

2 Trust, expectation, and frustration

In the process of their socialisation, humans tend to develop certain expectations, and expectation expectations (Luhmann 1995: 303-310) – i.e., expectations concerning other agents' expectations – based on what their life experiences are. They adapt to local regularities in that they start expecting certain things to happen or at least more or less count on the likelihood, that, what they think is probable to happen, will probably happen (cp., e.g., Millikan 2004; Mumford 2010; Poljanšek 2017; Rey et al. 2019; Rosenberg 2012; Williams 2019). These assumed probabilities refer to bygone occurrences and the way they are developed is described in different ways by different scientific branches.

Psychological theories state that human memory may be 'saved' schematically (for example, in *Memory Organisation Packets*; see, e.g., Schank 1980) and

[2] To live a good life, we assume, it may be helpful to manage one's own emotions in relation to their usefulness in interaction and in themselves for oneself, i.e., as a means to an interpersonal, and to a personal end. The discussion of aspects related to using sociosensitive/-active systems can also be understood as a reflection on potential interferences with or amplifications of living a 'good life'.

that iterated experience of situations leads to *scripts* (Schank/Abelson 1977) that are subsequently used to estimate what situation a person finds themselves in, what to expect from it, and how to behave. If the expected sequence of events, however the expectation may have been acquainted, does not actually take place, a person might react irritated or disoriented, and subsequently might feel frustrated, sad, angry – or, curious, surprised, excited. For example, a verbal or bodily expression of frustration might be observable in a person when they realise the train they've boarded does not arrive at the destination at the expected time of arrival. In this sense, irritation occurs when a person has started to expect a certain course of events that then takes another, unexpected trajectory. A then needed orienting response might be accompanied by subtle bodily behaviours (Bradley et al. 2012), or, even in infants, by surprise when a physically implausible event occurs (Bermúdez 2003: 54-55). From the presence of such reactions, we may infer that a person has had certain expectations.

Expectations are linked to the notion of trust.[3] For example, I tacitly 'trust' in the 'fact' that the sun will rise in the morning and that, under usual circumstances (i.e., in the environment that I am used to), an object will fall to the ground instead of starting to float in the air. "Trust" here refers to the assumption that my predictions will be correct with a certain probability that I infer from an observation of my environment. It is partly dependent on what I choose to believe and what I think to be justified to believe. We might call this *empirical trust.*[4] A social aspect comes into play when I assume that when I ask for a croissant in a bakery, I will most likely not receive a plush dinosaur instead. In other words, I 'trust' not only in the relative stability of certain

[3] Trust is conceived of as an important aspect of interpersonal relationships (Larzelere/Huston 1980) and of a functional society (Cook 2001), although 'more trust' is not necessarily always desirable (cp. Schelling 1984, p. 211; Goel et al. 2005). With regard to emotional bonds with technical objects "trust" is a well-researched topic in human-machine interaction studies (Cominelli et al. 2021; Hurlburt 2017; cp. Khavas 2021; Langer et al. 2019). For a distinction between reliability related to the so-called evidential view' with regard to *trust having reasons*, and *trust as a reason in itself*, related to the so-called 'assurance view', see Kaminski 2017.

[4] In contrast to *normative trust*, which involves desired outcomes, and predictions built on the grounds of believing that another agent should and will act in line with what I interpret as generally held values or commitments. For a discussion of trust types and the related distinctions between trust and reliability, as well as between empirically acquired expectations and normative expectations based on values, cp. Kaminski 2017.

physical circumstances, but also in the relative stability of societal, interactional, and linguistic systems and patterns, based on my experiencing them in the past. For this, I do not necessarily trust in the sense of ascribing another agent moral commitment – the baker does not need to commit to any moral beliefs to hand over a croissant or to use language in a common way. When I assume that he will not shoot me instead of selling me a croissant, I can, but do not need to ground my assumption in the belief that he is a moral agent, I can also more or less expect this based on experience (contextual world knowledge) and inferred probability.[5] In this sense, trust in reoccurring situational circumstances enables humans to get to know and to estimate what is to be expected with respect to different environments, items/artefacts, and other living beings' behaviour. As German Sociologist Niklas Luhmann puts it, trust is the only option besides "chaos and paralysing fear" (Luhmann 1979: 4), and "a complete absence of trust would prevent [a person] even from getting up in the morning" (Luhmann 1979: 4), because they wouldn't even count on, let's say, gravity. In this sense, expectations and expectation expectations make it easier for living beings to orient themselves and to navigate a complex physical world, as well as a society including interaction with other living beings. However, expectations and empirical trust do not need to be static. New experiences may lead to new or updated expectations.

Trust can be defined in several ways (for trust types cp. Müller 2009, p. 161-171) and through highlighting its connections to several concepts, such as decision-making (cp. Taddeo 2011), cooperation (Gambetta 1988), risk assessment (Siegrist 2021), or predictability (Tyler 2001, pp. 287-288; Reinhardt et al. 2017). Trust can be mapped to individual or group expectation differences, for example, some agents may be trusting less due to a fear of being exploited (Irwin et al. 2015). Other 'trust dimensions' in human-human interaction include for example *epistemic trust* (McCraw 2015; Sperber et al. 2010) which is built on the assumption that an interaction partner's information output is reliable. If an agent has reason to believe that others are 'unreliable narrators' they will be more cautious in trusting others' information (Fonagy et al. 2017). However, concerning the relation of reliability and predictability, being unreliable can, but does not need to be, interpreted as being unpredictable: *we can*

[5] In a universe where bakers tend to shoot their clients instead of selling croissants, I would expect otherwise. For the situational and contextual embeddedness of human behaviour cp. Bellon et al. 2022a; 2022b.

predict that someone will be unreliable. On another note, and with regard to assumed moral commitment, in human-human interaction trust is not always based on reliability, but seems to be gifted to agents who, for example, suggest holding deontological moral intuitions, such as *killing is always bad* (Everett et al. 2016). In this sense, trusting other autonomous agents comes with a "willingness to be in someone else's hands" and "living with trust involves profound vulnerability and some helplessness, which may easily be deflected into anger" (Nussbaum 2016: 94).

If things go differently than one would have thought (i.e., cognitively expected) or wanted (i.e., normatively expected) depending on the scale and quality the expectations implicit in our trusting have not been met, and under the condition that this deviation has been deemed negative, frustrated feelings can include the mentioned anger, but also, for example, bewilderment, grief, disappointment, resignation, powerlessness, helplessness, impotency, reluctance, and more. Less gravely, frustration may be expressed in a short-lived embodied moment of affective reflex to some unenjoyable sensation. For example, when catching a toe on a chair a person may raise a fist to the sky or use swear words.

To further clarify and link expectation to frustration, we may turn to Luhmann's distinction between cognitive and normative expectation. Arguably, graver feelings of frustration may more often be connected to normative expectations, while a mere moment of affective reflex may more often be connected to a cognitive expectation.However, the two ways of expecting can not only be distinguished by bodily reactions, but, according to Luhmann, by a subject's mental reaction to disappointment of expectations (see Luhmann 2014: 32-33). When a person *cognitively expects* a certain event to occur and it doesn't, they may be surprised, but will adjust their expectation and possibly change the script associated with the object or situation in question. They learn that they may have to expect differently in the future – i.e., they learn to predict more accurately. When, on the other hand, a person *expects normatively*, they react to unexpected events by holding on to their expectation and attribute the expectation disappointment as a fault or an error to the system that disappointed their expectation. Thus, those who expect normatively believe that they have some kind of normative claim to the fulfilment of the expectation in question. The normative expectation of certain behaviours seems to be appropriate only in cases where we have reason to believe that it is indeed, or ought to be, the function or task of the entity or agent in question to behave according to our expectation, and where someone – either the entity itself, someone else, or even a network of several addressees – can be held accountable if the normative expectation is disappointed. This is why, even when an actor confronted with a disappointment in their normative expectation changes the associated scripts, i.e., when they begin to cognitively expect something different, they may still hold on to the normative expectation that things *should* be different, and that the agent(s) assumed to be responsible *ought* to change them.

With this distinction in mind, to talk about justified or unjustified expectation relates to either an empirical/statistical or a moral take on human-world interaction, i.e., what (probably) is and what ought to be. What ought to be according to someone is possibly what could be; however, it might just as well be what cannot be (cp. for more detail Gransche 2022; Hubig 2006). Nevertheless, what is, has been, or is imagined to possibly be, no matter if cognitively or normatively expected, can be met with protest. We said above that reliability and predictability are not the same and 'we can predict that someone will be unreliable'; this can now be differentiated further: we can predict that someone will be unreliable, including the *cognitive expectation* that they will not do what we think they should, and the *normative expectation* that they nonetheless should.

2.1 Frustration functionality and contingency

From a point of view of a theory of society and of action, Hellmann (1994) defines a problem as the disappointment of an expectation. While the assumption that 'under normal circumstances everything will be approximately the way it used to be', enables the frictionless execution of commonplace processes like buying a bread in a bakery, having a conversation, or, being in a relationship involving emotional bonding, where such expectations are disappointed, according to this view, problems occur. Specifically, *social problems* occur when the disappointed interaction partner ascribes their feeling of disappointment to an act of another party's decision-making (ibid: 146). This may lead to *protest* (expressed on a scale of friendly request to spurs of anger and violent conflict), and demanding the decision be taken back, involving (1) the (potentially irrational) belief that the other agent has agency to do so – i.e., the other's manifestation *is perceived as an action*; and (2) a judgement of this action – or, if we subtract the imputed agency: of this *situation*, as undesirable; i.e., in short: "Social problems are what people think they are" (Fuller/Myers 1941: 25), according to what they think ought to be.

Now, the solution to such a problem can either be reached through expectation adjustment, for example, by challenging one's own perception and predicting (cognitive adaption), changing one's own evaluation of something as worthy of conflict (normative adaption), or, by maintaining the (cognitive and/or normative) expectation and trying to change the dynamics in upcoming iterations of formerly disappointing social encounters according to one's wishes (trying to enforce change). The latter only seems logical, where there is potential to actually change, i.e., where the "real-possible" is "receptive to being true" or the "potentially possible" is "receptive to being receptive to being true" (Gransche 2022: 67)[6]. In other words: at first glance and from a point of view of modal logic, protest only seems useful where change is potentially or "real-possible" (Gransche 2022).

However, even where change (a) is impossible, from a psychological point of view, protesting may have signalling and intrinsic value: even if the addressed agent does not actually have the agency to change their behaviour or state of existence, protest may (b) signal that the person expressing their frustration is an autonomous being, and (c) signal to oneself or others that 'things are wrong' and 'we should do something about that' (cp. Nussbaum 2016). Moreover, it may (d) be cathartic (cp. catharsis value in Opp 2019, *abreaction* in psychoanalysis, NeuroAffective Relational Model™ in psychotherapy). Regarding situations in which the addressed agent *is* autonomous, *has* agency, and their behaviour *can* be changed (a), the expression of frustration reactions has central social functions (see e.g., Planalp et al. 2006: chapter 20; Bartneck et al. 2020: 115-118). According to philosopher Victoria McGeer (2015), emotional expression can be described as a part of intersubjective *mind-shaping*. By expressing irritation or frustration people may indicate to one another that their normative expectations have been disappointed and they may implicitly or explicitly aim at changing the interactional scripts of interaction partners, thereby increasing the likelihood that others will behave according to their own expectations in the future. From such a perspective, mind-shaping could be described as a *reciprocal calibration or reciprocal recoding of the interactional scripts of autonomous systems towards their respective expectations through irritation (including a,b, c, and potentially d)*. Frustration expression is thus part of the fine mechanics of social attunement, insofar as it aims at changing the behavioural structure and expectations of other agents. Moreover, subjects

[6] Referring to Hubig 2006 referring to Zeno as cited by Diogenes Laertius 1972: 7.1 75-76.

signal to others that they themselves are autonomous agents with their own interests and needs.

Mindshaping as a social practice can be considered an attempt to deal with a fundamental socio-anthropological state of affairs, which Luhmann describes as *double contingency*. In contrast to the "simple contingency in the field of awareness" (Luhmann 2014: 26), within which things can always develop differently than a subject expects, the phenomenon of double contingency denotes the fact that subjects also have to deal with other subjects "who come into my range of vision as an ego-like source of original experience and action, as 'alter ego'" (Luhmann 2014: 25). The perception of an alter ego differs in terms of its level of complexity from the perception of, say, a stone, insofar as the ego perceiving the alter ego has to expect that the alter ego has its own expectations which, in turn, might themselves concern the expectations of the ego and *vice versa*.

Thus, when subjects interact with each other, they might reciprocally form *expectations of expectations*, expectations concerning the expectations of another agent. To close the circle and come back to the beginning of the chapter: Uncertainty about what to expect from another agent in cases of double contingency in human-human interaction is usually bridged by the phenomenon of trust. When confronted with another agent, we cannot know how they will behave in the future, but we can (tacitly or explicitly) choose to believe that they will behave in a certain way. Without *choosing* to do so,[7] what we do is not trust, but merely hope. By trusting we thus reduce the complexity of an unknown yet imagined future:

> "[R]ather than being just an inference from the past, trust goes beyond the information it receives and risks defining the future. The complexity of the future world is reduced by the act of trust. In trusting, one engages in action as though there were only certain possibilities in the future" (Luhmann 1979: 20).

However, trusting always remains risky for the trustor, insofar as they might have erred in believing in a course of upcoming events. A trustor knows that

[7] Logically, this holds true, no matter the extent to which one has chosen consciously – what we choose to believe may be subject to a preconscious choice, which remains a choice nonetheless, insofar as there is a modality of it possibly having been different. The psychological question is to what extent it is possible to enter the modal sphere we need to enter to change even preconscious choices.

the trustee could principally also behave differently from the way the trustor predicts. Thus, "trust reflects contingency" (Luhmann 1979: 24) and depends on contingency in the sense that the alter ego's ability to act differently than expected is a necessary condition to being able to trust it. To trust another person is not to rely on them as a mechanism that simply acts according to one's own expectations (for an elaboration of this point, cp. Kaminski 2017; Lahno 2002). Trusting another person means choosing to believe that they will act according to one's expectations, while simultaneously acknowledging that they are an autonomous source of original experiences and actions and could do otherwise. In other words: a human is the "animal that is allowed to make promises" (Nietzsche 1998: 35), i.e., to govern their own behaviour. Conversely, an alter ego that cannot help but act as we expect it to act can, in systematic terms, never be someone who can be trusted. From a perspective of potential conflict, pragmatically, besides from the potential signalling (b) and cathartic (d) benefits mentioned above, it would not only be futile to express frustration towards an agent that cannot make autonomous decisions, virtue ethicists may even argue that in certain situations, violent or stark protest may result in character damage in the agent performing it themselves (cp. Coeckelbergh 2021).

2.1.1 Frustration communication functionality

To sum up potential benefits of communicating frustration in human-human interaction:

a) *Changing the situation*: Communicating frustration may change the present situation (a_1) or even bring forth an altered iteration of a certain interaction in the future which will align more with the communicator's needs, wishes, and expectations (a_2).
b) *Signalling autonomy*: Communicating frustration signals to other agents that one is an autonomous being with needs, wishes, and expectations (b_1). Depending on how it is expressed and arguably, it *may* also signal that the person communicating frustrated feelings assumes their counterpart to be able to respond, thereby acknowledging their counterpart's autonomy (b_2).
c) S*ocial learning*: Communicating frustration signals to others, for example, what values are being held, which scripts are taken to exist or exist in

a person, group, or culture, what hurts another being, the presence of a potential danger, etc., and may thereby enable social learning.

d) *Catharsis-hypothesis*: Communicating frustration 'frees' a living being from being in a mentally and/or physically unpleasant state.

3 Expressing frustration towards non-living beings

3.1 Cognitive and normative expectations in dealing with non-living entities

> "We [...] are so completely blinded in our frustrations that sometimes if we have a sponge or (a piece of) wool in our hands we lift it up and throw it, as if we would thereby accomplish anything. [...] Often in this kind of blindness we bite the keys and beat against the doors when they are not quickly opened, and if we stumble on a stone we take punitive measures, breaking it and throwing it somewhere, and all the while we use the strangest language. [...] From such actions a person would get a notion of the irrationality in the affections and would perceive how we are blinded on such occasions, as though we were no longer the same persons who had earlier engaged in philosophical conversations." (Chrysippus: On the Affections, as referenced by Galen 1981: 280f.)

Although the dimension of 'real' consideration[8] of a living being's needs, wishes, and expectations may be missing in a non-living entity, we may still ask in what sense communicating frustration vis-à-vis a technical object or any other non-living entity might be useful to a living agent with regard to leading a 'good life'.

To that end, first of all, we might want to differentiate between expressing frustration with or without responsive other agents present. If you've caught your toe on a chair, communicating your frustration with the chair to the chair will not change the chair's behaviour in the future – in this sense you are not entering an interaction with the chair when you catch your toe on it and yell at it subsequently. Thus, here, (a) is not the case, for you are not even entering

[8] Real consideration would involve full-blown acknowledgement of another agent as a valuable being, cp. Bellon/Nähr-Wagener 2022.

*inter*action. Instead of *inter*acting, you may, however, *take action*, for example, put the chair somewhere else or heighten your attention when passing it. However, this will not be the result of having communicated your frustration with yourself to yourself and may still happen if you do not communicate your frustration – as long as you can feel your frustration without communicating it. Even if your action was the result of a 'self-communication', you would have entered an interaction with yourself, not with the chair. However, it might still be useful to express your frustration by yelling or gesturing angrily at the chair as it *might* alleviate some of your bodily stress (d), although others may argue it may even increase it. As with the chair, which is not an interaction partner (it does not take action itself as a reaction to your actions), an emotional reaction of expressed disappointment towards any other inanimate non-responsive object, even if useful in its potential to signal to other living beings that you are an agent (b_1), or, as a signal of danger (c), will remain, beyond its potential cathartic effect (d), inconsequential with regard to (a): Inanimate non-responsive objects will not change their behaviour if you express frustration, while living beings might.

This relates to the distinction between cognitive and normative expectation as follows: If a person expects a certain inanimate object to be relatively lightweight, but when lifting the object, it turns out to be quite heavy, the person will probably change their expectation or prediction concerning this specific object. The person had cognitively expected the weight of the object and shows a willingness to change this expectation if disappointed. If a stone is heavier than we thought, we probably would neither seriously blame it for its being heavier than expected, nor start looking for the accountable person behind this phenomenon to attribute responsibility to (after all, this is probably ourselves, as in: we've expected wrongly), nor attribute the deviation from our prediction to some autonomous will – i.e., we will usually not seriously normatively expect the stone to change its weight in the future. Normative expectation seems appropriate only in cases where one is dealing, either directly or indirectly, with double contingency, i.e., where one is dealing with other autonomous agents.[9] Relating to inanimate objects such as a stone or an artificial system is not usually a situation of double contingency.

However, inanimate objects differ in that they can be naturally given, cultivated, or produced. Most cultivation and all production originate from au-

[9] Which does not mean that people will not sometimes normatively expect "inappropriately", as in 'tilting at windmills'.

tonomous agents' decision-making. In this sense, "[w]hat resides in the machines is human reality, human gesture fixed and crystallized into working structures" (Simondon 2017: 18). In contrast to the case of a stone, concerning a produced entity, it might be possible to identify a correct addressee to attribute moral, legal, or political responsibility or, at least, distributed accountability for a product's performance and its consequences (cp. the concept of *responsibility networks* as laid out for example by Loh 2019 following Neuhäuser 2014). Correct addressees could be, for example, the company producing the product or its CEO, programmers, designers, whoever decides to purchase, install and use the technical system in any given context, as well as whoever will subsequently act according to an installed system's suggestions. Protesting undesired matters or effects thus might still be useful if addressed to these actors. Specifically, regarding artificial systems, we are justified in normatively expecting the reliable performance of certain functions, as the system was essentially determined by and built for the fulfilment of these functions and may have been purchased or installed to perform exactly that. The belief that this technical object *should* perform in a certain way, and that it is malfunctioning if it does not is justified by contract with an accountable company. Other normative expectations such as that production should not involve child labour or other exploitative measures may not be promised by contract but can still be addressed as requirements to an accountable party. However, where there is no address to attribute accountabilityto, i.e., where we are not dealing with the produced, *even though we might*, we are not *justified* in expecting normatively. For example, when a wooden stick used as a hiking rod breaks in two, we may feel frustrated, but will not have any entity addressable to normatively expect to receive a replacement or to attribute the frustration of our normative expectation to – we may raise our fist to the sky (with the possible benefits of b1, c, and d). But that will be inconsequential regarding (a).

Regarding cognitive expectations, a person might, no matter if dealing with naturally given or produced entities, willingly or without being aware of it, change their own behaviour in order to get the desired results, be it by learning to deal with some material, such as wood or clay, better, be it by adapting their own behaviour to match a technical system's abilities. For example, when a voice-controlled device does not understand a command, a person might normatively expect that the device *should* understand them better, but may still adapt their language commands so the system will understand them - the person will change their behaviour to get the cognitively expected results (while possibly still holding on to the normative belief that the system should be better and trying to ameliorate it).If a person insults a voice-controlled device (i.e., expresses their frustration), when it does not understand the command, we could argue that the expression of frustration is a result of invalid anthropomorphisation and an invalid attribution of agency, but it could just as well result from the mere effort it takes to have to leave a path of *cognitive expectation* that has already been trodden (one is used to use language a certain way and expects that to 'work'), i.e., to 'change scripts' or adapt habits..

3.2 Social interaction complexity levels and arguments for and against the use of socioactive and sociosensitive systems

Even if in the above example, the person insulting their voice-controlled device did not anthropomorphise (or zoomorphise) the inanimate entity, humans generally tend to do so (see, e.g., Marquardt 2017; McCarthy 1983; Picard 2008; Reeves/Nass 1996). It seems plausible to assume that with increasing complexity of interactional capability, users might contrafactually, but increasingly feel that they are dealing with situations of double contingency.[10] Although we do not necessarily need to ascribe double contingency or any humanlike characteristics to objects or events in order to have an emotional reaction towards them, and although social relations are usually defined by interacting with agents that are characterised as either having agency, or being a living individual organism (Radcliffe-Brown 1940: 2), if a technical system *seems* to react to our actions autonomously, we might still *feel* like we enter social relations with them. Crossley defines social relations as "lived trajectories of iterated interaction" (Crossley 2011: 28) between "actors", where to call something an actor implies "that it has a point of view and that this point of view matters and should be taken into account. It implies that the actor has a stake in the world under investigation, that it is meaningful for and matters to them." (Crossley 2011: 45)

However, humans may *experience* the relation with an inanimate object as a social relation in the sense of "lived trajectories of iterated interaction" re-

[10] This could be explained with Daniel Dennett's idea that in cases where the behaviour of an entity is too complex to predict with reference to its physical constitution, we may adopt what he calls the "intentional stance" which predicts behaviour by attributing desires and beliefs (cp. Dennett 1971; Dennett 2009).

gardless of whether the other (shm)agents[11] really have or merely simulate having a stake in the world. While this includes the possibility to feel like having social relations even with non-living entities such as models or algorithms (Lange 2021: 120), with artificial systems succeeding in realizing sociosensitive and socioactive capacities – i.e., capacities to identify social facts or cues and to process them in a way that will alter their own output in a way that takes into account the identified social cues – this feeling might become stronger. Sociosensitive and -active systems might register users' emotional frustration or disappointment and respond to it by modifying their behaviour accordingly, thereby giving the impression of interactional social relations – and changing the way users interact: *For example, frustration expression might suddenly become functional in the above mentioned sense of (a) when a system reacts to it.* A system may also simulate having its own needs, wishes, and expectation expectations, thereby giving the impression of interactional double contingency.

State-of-the-art technical emotion recognition systems can already detect social signals (Vinciarelli 2017) and infer emotions and even personality traits from such signals, for example, by voice analysis (Deng et al. 2017; Sagha et al. 2017; Schroder et al. 2015). They can be controlled by gestures (Obaid et al. 2014), so that it is possible to make them react to emotions inferred from body posture, movement, or, other cues such as body temperature, etc. In addition, socioactive technical systems are already designed to display signs of emotion that humans can interpret and understand (Breazeal 2004; Nitsch/Popp 2014; Salem/Dautenhahn 2017).

Hypothetically, there are at least three different levels, at which emotion- or sociosensitive and -active systems might be able to react or respond to the frustration of interaction expectations of its users. They may be able to take into account humans' disappointment by either I) exhibiting some sort of *recognition behaviour* when its users show signs of disappointment, II) by *switching to other behavioural sequences pre-coded in the system*, or III) by using some sort of adaptive learning mechanism trying to *find new and more accepted behaviour sequences*. Examples of existing systems can be found for level I and II, while level III has not yet been realised in the sense we will lay out.

[11] David Enoch (2006) calls a *shmagent* an agentlike non-agent performing *shmactions*, i.e., actionlike non-actions.

I. On the first level, technical systems might show some sort of recognition of the disappointment of its users without, however, changing their behavioural patterns. A system could mimically, symbolically, verbally, or through movement express that it registers a user's disappointment. This could have the simple advantage of making the user feel seen and acknowledged in their frustration. From a systems design perspective, the advantage could be a resulting mitigation of user frustration, which might otherwise have been directed at the technical object or have led to an interaction termination. User satisfaction could increase and result in prolonged interaction and heightened willingness to perform actions suggested by the system. Users might even feel a sense of 'respect' generated by the machine if their emotion expression is met with a 'reaction'. If respectful interaction is defined as an interaction, in which the mere feeling of being respected is the measure for respectful behaviour (Quaquebeke/Eckloff 2010), regardless of whether the interaction partner has actually been respected or not, one could argue that such an interaction might be desirable in a kind of reversed sense of (b2): A user might feel acknowledged as an autonomous being. Recognition could also lead to an increase in users' awareness of their own emotional states. If the system were able to accurately interpret markers of a subject's emotional states, it could help users to distinguish their own emotional reactions that they may not have registered without the help of this recognition. — If the laid out interactional consequences are deemed desirable to live a good life, expressing frustration towards a 'level I' system might be useful in this regard. On the other hand, if, for example a ticket machine detects users' frustration with the machine when the process of buying the ticket takes too long and the train is leaving, wa soothing, but unnecessarily time-consuming 'I see you'-performance may in this case lead to even greater frustration. After all, to live a good life, we may just want the machine to do its job and not complicate things for us by pretending to have humanlike qualities.

II. On the second level, sociosensitive and -active systems might additionally have the ability to switch between different behavioural scripts with respect to different types of frustration expressions of its users. For example, an artificial pet could have different behavioural scripts regarding its response to petting. If a user reacts with frustration to the artificial pet reacting not euphorically enough, or, too much, to the petting, the pet could adapt its behavioural script accordingly. Of course, the problem of exactly how the system is supposed to recognize what its user's frustration refers to in a particular case would need to be solved. For example, the system could pose a question and offer several behaviour options the user could then choose from. Users who express frustration may not only feel level-I acknowledged, they may also be more satisfied with the felt interaction quality, as the taking into account their wishes by providing options in case of frustration seems more interactive than just acknowledging user's frustration. With such a system, frustration expression would lead to (a1): changing the situation, and potentially to (a2): ameliorated quality of upcoming interaction iteration, if the system processes stable user preferences.

III. On a third level, one could imagine that technical objects could be additionally endowed with the ability to dynamically adapt to users' emotions through a kind of adaptive and associative learning mechanism. Instead of merely registering a user's emotional frustration (level I) and offering options in case of frustration (level II), such a system would be able to recognize emotional disappointments of a user, speculatively infer the user's expectations underlying such disappointments, and adjust its behaviour accordingly, creatively, and, quite randomly, for example, by accessing information from other services and by trial and error of applying the information. Let us give a highly speculative, potentially dangerous scenario: A system detects its users disappointment and tries to mitigate user frustration by accessing, for example, a database in which reddit commentary has been annotated with hints and tips on how to mitigate frustration in relationships. The system might be able to semantically extract the information that buying disappointed people flowers may lessen their frustration. It then may order flowers online and have them sent to the user with a note saying sorry. As a system will arguably never be able to produce its own creative solutions that are not based on any given data accessible to it (i.e., datafied information), one problem of course is the missing capacity of reason in a machine. While a human being might know intuitively that insulting a friend will not alleviate the friend's frustration in a conflict situation, if the data base holds this information, the system has no capacity to reflect on that (cp. Neff/Nagy 2016). Nonetheless, let us imagine a self-adapting system as a 'wish machine', learning to get to know its users and their specific preferences and idiosyncrasies way beyond what we know as personalized human-technology interaction: with such an idealized, as well as a with a more realistic, yet still surprisingly 'attentive' and adaptive system, users might get the impression that they

> are dealing with an almost empathetic counterpart, with whom they can interact smoothly, and who might even seem to engage in social learning (c) or overall mindshaping. Frustration expression towards such a system may lead to somewhat unexpected, surprising, or even random results, which subsequently might lead to users feeling 'scammed', or, potentially a lot closer to being in a relation of double contingency.

With regard to the question of a 'good life', we may add the following problems that may arise, aside from the usual, such as data privacy, data trading, potential manipulation of users through micro-targeting, and perpetuated unethical takes stemming from training data or other accessed and processed information, .

Firstly, continuous interaction with systems tuning in to the assumed expectations of their users more and more – and hypothetically reaching the idealised successful level III version of the wish machine[12] – may have the effect that human agents may start to expect the same frictionless execution of fulfilling their desires in human-human interaction as well (cp. Bisconti 2021). In other words: Frequent interactions with artificial systems that align themselves with the expectations and needs of their users could, in the long run, lead users to normatively expect interaction partner's not to have their own stakes in the world. Or, at least, they may tend to become frustrated more quickly in interactions with other actors if they, in turn, have normative expectations that do not directly align with their own normative expectations. To equip a system with the capacity to refuse to perform the function its user intended it to perform, or, to equip it even with the capacity to demand justification from its user for the user's actions would counteract this problem. However, this kind of simulation of double contingency may seem unethical to some who may argue that a system – which by definition cannot be autonomous in the sense of having the intrinsic will or desire to follow its own needs or to consider others' – should not by any means signal to us that it actually might have those capabilities or encourage us, for example by design, into thinking so. The implementation of such simulation could also be rejected on the grounds of arguing that the more an artificial system pretends

[12] Which would come with its own set of new operational and ethical problems, such as how the system would infer your wishes from the collected cues, and, if it should follow what it infers from that data as your wishes or have a function to deny functions based on ethical consideration or implemented rules.

to be an actual alter ego, the more likely its users are inclined to experience this system as an "ego-like source of original experience and action" (Luhmann 1979: 25), which could result in users having highly unrealistic expectations towards such technical systems. They might thus, for example, tend to overtrust them (Robinette et al. 2016), become frustrated with the system more easily and proportionally to the growth of the gap between the expected and the real-possible capacities of the system (cp. van den Berg 2011), or start to see the machine as an alter ego more worthy of care than actual alter egos in their environment. Furthermore, not only might such a system provide us with a poor paradigm for interpersonal interactions that might lead us to normatively expect that other subjects should not develop or express normative expectations toward us. The converse is also true: the meaning of trust and promise can only be learned in concrete confrontation with situations of genuine double contingency, only in confrontation with the other's freedom, as well as the other's confrontation with one's own freedom.

Moreover, the fact that users might experience such artificial systems as alter egos could increasingly obscure the fact that a technical object itself is neither responsible for the functions it was constructed to perform, nor for the way in which it fulfils these functions. If artificial systems give users the impression that they are autonomously acting entities with their own needs and stakes in the world, this could reinforce the perception that these machines also bear responsibility for their behaviour. In an extreme case, one could imagine that manufacturers or other stakeholders of such machines may try to use this circumstance to conceal their own responsibility behind the fact that humans increasingly perceive the sociosensitive and -active machines as alter egos, as independent sources and potential addressees for the attribution of responsibility. Expressing frustration with a technical object, or even fighting with it, may in such cases result in a person losing sight of the actual addressees of the attribution of responsibility. If there is an accountable person behind a certain operation a technical object performs, it might be advisable not to draw attention away from that fact, or, at least, bring it back to attention frequently.

Another argument against the simulation of agency and double contingency is that we might just want a machine to do what we want it to do. If the ticket machine suddenly starts refusing to sell us tickets, we may say that it has lost its value to us. If systems as complex as we have imagined with the third level ever exist, users may need to decide for themselves what settings they will prefer: simulation of agency by denying to perform certain functions, or, 'obedient' function-fulfilling machines that 'just do the work' (cp. for conflicting imaginaries in an application field for such a scenario for example Depunti 2022). Last but not least, operational problems might occur, for example when emotion recognition does not work well enough or on false premises.

4 Conclusion

With respect to the chapter's initial question of whether one should fight with a machine, we conclude: Where the expression of a frustration is an end in itself and serves the human need to release tension (d), where it does at the same time not damage the frustrated person's character (Coeckelbergh 2021) or any other entity (cp. Cosio/Taylor 1992) and does not signal to others that we can and should be ruthless with the world's material, it might be useful to express this frustration towards a machine. Where the expression is meant as a means to an end, that is, to change the situation (a_1) or invoke an altered future in which the interaction should take place otherwise (a_2), there are two possibilities: you are either dealing with a technical object that has no capacity to react to your wishes or you are dealing with a technical object that has a limited capacity to take into account your wishes and to alter their own behaviour. Thus, implementing sociosensitive and socioactive capabilities into technical objects changes the usefulness of communicating frustration vis-à-vis the system, therebypotentially adding a new layer to the way living beings relate to non-living entities. With regard to (b_1) there might be situations in which signalling your autonomy by protesting against inanimate objects (such as institutions and norms, which may be embodied by technical objects as well, cp. Winner 1980) may be useful. Concerning (c) bidirectional 'mind-shaping' through frustration expression may be metaphorically possible in highly adaptive, learning systems modifying their output taking into account individual, group, or cultural preferences, and for users that form habits from frequent interaction with systems. It is unclear whether this is desirable or not.

As we have seen, sociosensitive and -active technical objects or systems may be able to somewhat take into account humans' frustration. However, as they do not possess decision-making agency in a sense that allows for a 'true' consideration of other beings' emotional states, i.e., a system cannot fully acknowledge a human being as an autonomous being (cp. Siep 2022), it is to be decided on a case-to-case basis if it is desirable that a system will be designed in such a manner that it will be able to adjust its behaviour according to a user's inferredemotional states – and which ones. In any case the experience of useful frustration expression should not lead a user into trusting the system as they would trust an agent with full-blown agency and the capacity to reason and self-reflect: Trust in interpersonal relationships necessarily involves contingency, i.e., an interaction partner's possibility to act differently than expected, even if they continue to not disappoint the trust that is placed in them. An artificial system that adapts to the expectations of its users in order not to disappoint them in the further course of interaction is not an agent that can be 'trusted' in this sense. Thus, it's not that you can't trust machines because they might deceive you. It's that you can't trust a machine because it can't choose to act differently than expected.

5 Reflection and Outlook

Transferability of human-human interaction concepts to human-machine or human-technology relations is in itself highly problematic. On the one hand, a technical object can be understood as 'just another thing' in the sense of it being predictable or unpredictable just as much as a stone or the weather. We may try to understand the interconnected ways of its inner operations, dynamics, and its links to other entities and laws of local nature in the same way as we would with any other object. On the other hand, machine behaviour can be irritating to living beings in entirely new ways. While humans tend to be able to more or less predict what other humans may do, certain technical objects, such as (embodied) algorithms may be, although produced by humans, totally opaque to a human observer, and others may irritate human expectation by looking similar but acting different: For example, it may be quite predictable to human beings how other human beings drive their cars, what they mean by certain traffic-related gestures, they might even infer from a certain driving style if the driver is drunk, etc. When observing a so-called 'autonomous' car, these inferences are not valid anymore to predict the car's behaviour in the same way it would be probable with regard to a human-driven car. In this sense, technical objects 'behave' according to their own logic, which might be very unfamiliar to human expectations and, therefore, be quite unpredictable. Concerning these dimensions of technology, human observers have many new expectations to acquire and may be surprised or frustrated in their sedimented expectations more than a few times.[13] For this reason, some researchers call for a science of machine behaviour (Rahwan et al. 2019), or for so-called *mechanologists*, i.e., psychologists and sociologists of technical objects (Simondon 2017). From this perspective, phenomena in which systems show "behaviour that satisfies the literal specification of an objective without achieving the intended outcome" (Kraknova et al. 2020) are not only entertaining and interesting examples of machinic logic,[14] but just as much show how human agents expect and predict. Once we learn that our own expectations have their own human, or even individual logic and are just one possibility of many ways to be in the world, we might be more open to react to the unexpected – where it isn't existentially hurtful – with an extended interest in the otherness of the entity we weren't able to predict – and with unjudgmental surprise and curiosity. From there on, with or without expressing frustration, the (inter)action options, in cases where we are not forced into the relation, will still be the usual: love it, change it, and/or leave it.

References

Bartneck, Christoph/Belpaeme, Tony/Eyssel, Friederike/Kanda, Takayuki/Keijsers, Merel/Šabanović, Selma (2020): Human-robot interaction. An introduction, Cambridge, New York, NY, Port Melbourne: Cambridge University Press.

Bellon, J., Nähr-Wagener, S. (2022): "Einleitung". In: Jacqueline Bellon/Bruno Gransche/Sebastian Nähr-Wagener (eds.), Soziale Angemessenheit, Wiesbaden: Springer VS, pp. 33-47.

Bellon, Jacqueline/Eyssel, Friederike/Gransche, Bruno/Nähr-Wagener, Sebastian/Wullenkord, Ricarda (2022): Theory and Practice of sociosensitive and socioactive systems, Wiesbaden: Springer.

Bellon, Jacqueline/Gransche, Bruno/Nähr-Wagener, Sebastian (2022): Soziale Angemessenheit – Forschung zu Kulturtechniken des Verhaltens, Wiesbaden: Springer.

[13] See, e.g., https://twitter.com/llsethj/status/1512960943805841410?lang=de. Here a police officer tries to adjust their expectation when stopping a driverless car which does not react to language and procedures the officer is habituated to use in cases like this.

[14] For a list of examples from the field of *specification gaming* see: https://heystacks.com/doc/186/specification-gaming-examples-in-ai---master-list.

Bermúdez, José L. (2003): Thinking without words, Oxford: Oxford University Press.

Bisconti, Piercosma (2021): "How Robots' Unintentional Metacommunication Affects Human-Robot Interactions." In: Minds & Machines, pp. 487-504.

Bradley, Margaret M./Keil, Andreas/Lang, Peter J. (2012): "Orienting and emotional perception: facilitation, attenuation, and interference." In: Frontiers in Psychology 3, article 493.

Breazeal, Cynthia (2004): "Function Meets Style: Insights From Emotion Theory Applied to HRI." In: IEEE Transactions on Systems, Man and Cybernetics, Part C 34, pp. 187-194.

Brscić, Drazen/Kidokoro, Hiroyuki/Suehiro, Yoshitaka/Kanda, Takayuki (2015): "Escaping from Children's Abuse of Social Robots." In: Julie A. Adams/William Smart/Bilge Mutlu et al. (eds.), HRI'15. Proceedings of the 2015 ACM/IEEE International Conference on Human-Robot Interaction: March 2-5, 2015, Portland, OR, USA, [New York]: ACM, Associaton for Computing Machinery 2015, pp. 59-66.

Coeckelbergh, Mark (2021): "How to Use Virtue Ethics for Thinking About the Moral Standing of Social Robots: A Relational Interpretation in Terms of Practices, Habits, and Performance." In: International Journal of Social Robotics 13, pp. 31-40.

Cominelli, Lorenzo/Feri, Francesco/Garofalo, Roberto/Giannetti, Caterina/Meléndez-Jiménez, Miguel A./Greco, Alberto/Nardelli, Mimma/Scilingo, Enzo P./Kirchkamp, Oliver (2021): "Promises and trust in human-robot interaction." In: Scientific reports 11, 9687.

Cook, Karen (ed.) (2001): Trust in society, New York: Sage.

Cosio, Michael/Taylor, Gregg (1992): "Soda pop vending machine injuries. An update". In: Journal of Orthopaedic Trauma 6/2, pp. 186-189.

Crossley, Nick (2011): Towards Relational Sociology, New York: Routledge.

Deng, Jun/Eyben, Florian/Schuller, Bjorn/Burkhardt, Felix (2017): "Deep neural networks for anger detection from real life speech data." In: International W. o. C. B. A. Recognition (ed.), 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). 23-26 Oct. 2017, [Piscataway, NJ]: IEEE 2017, pp. 1-6.

Dennet, Daniel C. (2009): "Intentional Systems Theory." In: Ansgar Beckermann/Brian

Dennett, Daniel C. (1971): "Intentional Systems." In: Journal of Philosophy 68, pp. 87-106.

Depounti, Iliana/Saukko, Paula/Natale, Simone (2022): Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend, (tba).

Ellis, Ralph D./Zachar, Peter (eds.) (2012): Categorical versus dimensional models of affect. A seminar on the theories of Panksepp and Russell, Amsterdam, Philadelphia: John Benjamins Pub.

Everett, Jim A. C./Pizarro, David A./Crockett, Molly J. (2016): "Inference of trustworthiness from intuitive moral judgments." In: Journal of experimental psychology. General 145, pp. 772-787.

Fonagy, Peter/Luyten, Patrick/Allison, Elizabeth/Campbell, Chloe (2017): "What we have changed our minds about: Part 2. Borderline personality disorder, epistemic trust and the developmental significance of social communication." In: Borderline personality disorder and emotion dysregulation 4, 9.

Galen (1981^2): On the doctrines of Hippocrates and Plato, Berlin: Akademie.

Gambetta, Diego (1988) "Can We Trust Trust?" In: Diego Gambetta (ed.), Trust: Making and Breaking Cooperative Relations, Oxford: blackwell, pp. 213-237.

Goel, Sanjay/Bell, Geoffrey/Pierce, Jon (2005): "The Perils of Pollyanna: Development of the Over-Trust Construct." In: Journal of Business Ethics 58, pp. 203-218.

Gransche, Bruno (2022): "Ask what can be! Modal critique and design as drivers of accidence." In: Claudia Mareis/Moritz Greiner-Petter/Michael Renner (eds.), Critical by Design? Genealogies, Practices, Positions, Bielefeld: transcript, pp. 64-79.

Gunkel, David J. (2018): Robot rights, Cambridge, Massachusetts: MIT Press.

Hellmann, Kai-Uwe (1994): "Zur Eigendynamik sozialer Probleme." In: Soziale Probleme 5, pp. 144-167.

Hubig, Christoph (2006): Die Kunst des Möglichen I, Bielefeld: transcript.

Hurlburt, George (2017): "How Much to Trust Artificial Intelligence?" In: IT Professional 19, pp. 7-11.

Irwin, Kyle/Edwards, Kimberly/Tamburello, Jeffrey A. (2015): "Gender, trust and cooperation in environmental social dilemmas." In: Social science research 50, pp. 328-342.

Kaminski, Andreas (2017): "Hat Vertrauen Gründe oder ist Vertrauen ein Grund? – Eine (dialektische) Tugendtheorie von Vertrauen und Vertrauenswürdigkeit." In: Jens Kertscher/Jan Müller (eds.), Praxis und ›zweite Natur‹, Münster: mentis, pp. 167-188.

Khavas, Zahra R. (2021): A Review on Trust in Human-Robot Interaction 2021.

Kraknova, Victoria/Uesato, Jonathan/Mikulik, Vladimir/Rahtz, Matthew/Everitt, Tom/Kumar, Ramana/Kenton, Zac/Leike, Jan/Legg, Shane (2020): Specification gaming: the flip side of AI ingenuity; https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity.

Lahno, Bernd (2002): Der Begriff des Vertrauens, Paderborn: mentis.

Lange, Markus (2021): Affekt, Kalkulation und soziale Relation. Ungewissheitsarrangements der Finanzmarktpraxis, Wiesbaden: Springer.

Langer, Allison/Feingold-Polak, Ronit/Mueller, Oliver/Kellmeyer, Philipp/Levy-Tzedek, Shelly (2019): "Trust in socially assistive robots: Considerations for use in rehabilitation." In: Neuroscience & Biobehavioral Reviews 104, pp. 231-239.

Larzelere, Robert/Huston, Ted (1980): "The Dyadic Trust Scale: Toward Understanding Interpersonal Trust in Close Relationships." In: Journal of Marriage and Family 42, pp. 595-604.

Loh, Janina (2019): "Responsibility and Robot Ethics: A Critical Overview." In: Philosophies 4/58.

Luhmann, Niklas (1979): Trust and Power, Chichester: Wiley and Sons.

Luhmann, Niklas (1995): Social Systems, Stanford: SUP.

Luhmann, Niklas (2014): A Sociological Theory of Law, New York: Routledge.

Marquardt, Manuela (2017): Anthropomorphisierung in der Mensch-Roboter Interaktionsforschung: theoretische Zugänge und soziologisches Anschlusspotential, Berlin, Essen: Universität Duisburg.

Mayntz, Renate/Nedelmann, Birgitta (1987): "Eigendynamische soziale Prozesse. Anmerkungen zu einem analytischen Paradigma." In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 39, pp. 648-668.

McCarthy, John (1983): "The Little Thoughts of Thinking Machines." In: Psychology Today, pp. 46-49.

McCraw, Benjamin W. (2015): "The Nature of Epistemic Trust." In: Social Epistemology 29, pp. 413-430.

McGeer, Victoria (2015): "Mind-making practices: The social infrastructure of self-knowing agency and responsibility." In: Philosophical Explorations 18, pp. 259-281.

McLaughlin/Sven Walter (eds.), The Oxford Handbook of Philosophy of Mind.

Millikan, Ruth G. (2004): Varieties of meaning, Cambridge, Mass: MIT Press.

Müller, Jeanette (2009): Vertrauen und Kreativität, Frankfurt am Main: Lang.

Mumford, Stephen (2010): David Armstrong, Stocksfield UK: Acumen.

Neff, Gina/Nagy, Peter (2016): "Talking to Bots: Symbiotic Agency and the Case of Tay". In: Int. Journal of Communication 10, pp. 4915-4031.

Neuhäuser, Christian (2014): "Roboter und moralische Verantwortung." In: Eric Hilgendorf (ed.), Robotik im Kontext von Recht und Moral, Baden-Baden: nomos, pp. 269-286.

Nietzsche, Friedrich (1998): On the Genealogy of Morality, Indianapolis: Hackett.

Nitsch, Verena/Popp, M. (2014): "Emotions in robot psychology." In: Biological cybernetics 108, pp. 621-629.

Nussbaum, Martha (2016): Anger and Forgiveness, New York: OUP.

Obaid, Mohammad/Kistler, Felix/Häring, Markus/Bühling, René/André, Elisabeth (2014): "A Framework for User-Defined Body Gestures to Control a Humanoid Robot." In: International Journal of Social Robotics 6, pp. 383-396.

Opp, Karl-Dieter (2019): The Rationality of Political Protest, New York: Routledge.

Picard, Rosalind W. (2008): "Toward Machines With Emotional Intelligence." In: Gerald Matthews/Moshe Zeidner/Richard D. Roberts (eds.), The Science of Emotional Intelligence: Knowns and Unknowns, Oxford, England: Oxford University Press, pp. 396-416.

Planalp, Sally/Fitness, Julie/Fehr, Beverley: "Emotion in Theories of Close Relationships." In: Anita L. Vangelisti/Daniel Perlman (eds.), The Cambridge handbook of personal relationships, Cambridge: Cambridge University Press, pp. 369-384.

Poljanšek, Tom (2017): "Die Vorstrukturierung des Möglichen – Latenz und Technisierung." In: Alexander Friedrich/Petra Gehring/Christoph Hubig et al. (eds.), Technisches Nichtwissen. Jahrbuch Technikphilosophie, Baden-Baden: Nomos, pp. 15-40.

Quaquebeke, Niels van/Eckloff, Tilman (2010): "Defining Respectful Leadership: What It Is, How It Can Be Measured, and Another Glimpse at What It Is Related to." In: Journal of Business Ethics 91, pp. 343-358.

Radcliffe-Brown, A. R. (1940): "On Social Structure." In: The Journal of the Royal Anthropological Institute of Great Britain and Ireland 70, pp. 1-12.

Rahwan, Iyad/Cebrian, Manuel/Obradovich, Nick/Bongard, Josh/Bonnefon, Jean-François/Breazeal, Cynthia/Crandall, Jacob W./Christakis, Nicholas A./Couzin, Iain D./Jackson, Matthew O./Jennings, Nicholas R./Kamar, Ece/Kloumann, Isabel M./Larochelle, Hugo/Lazer, David/McElreath, Richard/Mislove, Alan/Parkes, David C./Pentland, Alex './Roberts, Margaret E./Shariff, Azim/Tenenbaum, Joshua B./Wellman, Michael (2019): "Machine behaviour." In: Nature 568, pp. 477-486.

Rammert, Werner/Schulz-Schaeffer, Ingo (2002): Technik und Handeln - wenn soziales Handeln sich auf menschliches Verhalten und technische Artefakte verteilt, Berlin: Technische Universität Berlin.

Reeves, Byron/Nass, Clifford (1996): The media equation. How people treat computers, television, and new media like real people and places, s.l.: CSLI Publications.

Reinhardt, Jakob/Pereira, Aaron/Beckert, Dario and Bengler, Klaus (2017): "Dominance and movement cues of robot motion: A user study on trust and predictability". In: IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1493-1498.

Rey, Arnaud/Minier, Laure/Malassis, Raphaëlle/Bogaerts, Louisa/Fagot, Joël (2019): "Regularity Extraction Across Species: Associative Learning Mechanisms Shared by Human and Non-Human Primates." In: Topics in cognitive science 11, pp. 573-586.

Robinette, Paul/Li, Wenchen/Allen, Robert/Howard, Ayanna M./Wagner, Alan R. (2016): "Overtrust of robots in emergency evacuation scenarios." In: Christoph Bartneck (ed.), The Eleventh ACM/IEEE International Conference on Human Robot Interaction, Piscataway, NJ: IEEE Press 2016, pp. 101-108.

Rosenberg, Alexander (2021): Philosophy of science. A contemporary introduction, New York, NY, London: Routledge.

Ryland, Helen (2021): "Could you hate a robot? And does it matter if you could?" In: AI & Society 36, pp. 637-649.

Sagha, Hesam/Deng, Jun/Schuller, Bjorn (2017): "The effect of personality trait, age, and gender on the performance of automatic speech valence recognition." In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), [Piscataway, New Jersey]: IEEE 2017, pp. 86-91.

Salem, Maha/Dautenhahn, Kerstin (2017): "Social Signal Processing in Social Robotics." In: Alessandro Vinciarelli/Judee K. Burgoon/Maja Pantic et al. (eds.), Social signal processing, Cambridge: Cambridge University Press, pp. 317-328.

Schank, Roger C. (1980): "Language and Memory." In: Cognitive Science 4, pp. 243-284.

Schank, Roger C./Abelson, Robert (1977): Scripts, plans, goals, and understanding: An inquiry into human knowledge structures, Hillsdale, NJ: Lawrence Erlbaum Associates.

Schelling, Thomas (1984): Choice and Consequence, Cambridge: Harvard University Press.

Schroder, Marc/Bevacqua, Elisabetta/Cowie, Roddy/Eyben, Florian/Gunes, Hatice/Heylen, Dirk/ter Maat, Mark/McKeown, Gary/Pammi, Sathish/Pantic, Maja/Pelachaud, Catherine/Schuller, Bjorn/Sevin, Etienne de/Valstar, Michel/Wollmer, Martin (2015): "Building autonomous sensitive artificial listeners." In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII 2015). Xi'an, China, 21 - 24 September 2015, Piscataway, NJ: IEEE 2015, pp. 456-462.

Siegrist, Michael (2021): "Trust and Risk Perception: A Critical Review of the Literature." In: Risk Analysis 41, pp. 480-490.

Siep, Ludwig (2022): "Angemessenheit und Anerkennung aus philosophischer und philosophiehistorischer Perspektive." In: Bellon, Jacqueline/Gransche, Bruno/Nähr-Wagener, Sebastian (eds.), Soziale Angemessenheit, Wiesbaden: Springer VS, pp. 49-64.

Simondon, Gilbert (2017): On the mode of existence of technical objects, Washington: Univocal Publishing.

Sperber, Dan./Clément, Fabrice/Heintz, Christoph/Mascaro, Olivier/Mercier, Hugo/Origgi, Gloria/Wilson, Deirdre (2010): "Epistemic Vigilance." In: Mind & Language 25, pp. 359-393.

Taddeo, Mariarosaria (2011): "Defining trust and e-trust: from old theories to new problems." In: Anabela Mesquita (ed.), Sociological and philosophical aspects of human interaction with technology, pp. 24-37.

Tyler, Tom R. (2001): "Why do people rely on others? Social identity and social aspects of trust." In: Karen S. Cook (ed.), Trust in society, New York: Sage, pp. 285-306.

van den Berg, Bibi (2011): "The Uncanny Valley Everywhere? On Privacy Perception and Expectation Management". In: Fischer-Hübner, Simone et al. (eds), Privacy and Identity Management for Life, Heidelberg: Springer, pp. 178-191.

Vinciarelli, Alessandro (2017): "Introduction: Social Signal Processing." In: Alessandro Vinciarelli/Judee K. Burgoon/Maja Pantic et al. (eds.), Social signal processing, Cambridge: Cambridge University Press, pp. 1-8.

Williams, Robert G. (2019): The metaphysics of representation, Kettering: Oxford University Press.

Winner, Langdon (1980): "Do Artifacts Have Politics?" In: Daedalus 109/1), pp. 121-136.

Empathic Machines?
Ethical Challenges of Affective Computing from a Sustainable Development Perspective

Cordula Brand, Leonie N. Bossert, Thomas Potthast

1 Introduction

The question whether machines can recognise and simulate emotions is currently researched and discussed intensely in science, industry, and the public.[1] The corresponding technology is called *Affective Computing* (AC). The concept and possible applications of AC gain much attention due to the high potential as well as risks for society such as the potential misuse of highly sensitive data on the one hand or fostering participation within society through sensitive technology on the other (Devillers 2021; Cowie 2015). AC is linked to two main goals in the field of machine learning. Firstly, machines should be enabled to *recognise emotional states* of people to adapt the machine's behaviour to these states, i.e., they should be made 'empathic'. Secondly, and especially in the case of conversational systems like chat bots, avatars, or robots, for example in the care sector, they should be able to *simulate emotions* convincingly to enrich and simplify human-computer interaction (HCI). For implementing both recognition and simulation, the following parameters are mainly used: facial expression, posture, gestures, and speech. Depending on the technical

[1] We thank the editors for their valuable feedback and Lukas Weber for assistance in preparing the manuscript. The paper has been developed in the project ("Orientation towards the common good in the digital age - Transformation narratives between planetary boundaries and artificial intelligence" (Gemeinwohlorientierung im Zeitalter der Digitalisierung. Transformationsnarrative zwischen Planetaren Grenzen und Künstlicher Intelligenz), funded by the German Environment Agency (Umweltbundesamt - UBA, FKZ 3718111050).

equipment, recognition may also include the possibility of collecting additional physiological data like skin temperature and skin conductivity (Picard 2000: chapter 2). Various possible applications are tested, developed, and discussed: from the gaming scene to sexbots, from automotive industry to advertisement and possibilities for self-optimization, and within the care- and education sector.

All of them come along with diverse ethical aspects, which need to be considered when discussing AC technologies. We argue in this chapter that the ongoing ethical considerations so far lack an important perspective: orientation towards a justice-based approach – we use the Sustainable Development framework –, which may help in developing and using AC technologies that will benefit all humans, not only privileged groups.

We shall first give an overview on the ethical issues that need to be addressed regarding AC. Some of them are rather general and have already been mentioned regularly (Hagendorff 2020), as they emerge in the overall context of Artificial Intelligence (AI)-technologies and Big Data, like protection against discrimination, equity of access and protection of the privacy of users and cybersecurity. Affective Computing, however, raises additional ethical issues. This technology seems to be able to change our understanding of what it means to be a human being more severely than other applications of AI.

In a second step we will thus summarize these anthropological concerns. Here, aspects are addressed that are rarely mentioned in AI-guidelines as well as in the academic discourse on AI, such as, for example, solidarity, inclusion, and diversity. These aspects are only slowly entering the academic as well as public debate concerning AI. A reason for this can be found in a technology-focussed ethical approach (rather than taking into account the human condition) that at the same is concentrating too much on an individual level. Some of these missing aspects, this is our argument, can be covered by the normative concept of Sustainable Development (World Commission on Environment and Development 1987), which is a justice-based approach.

Therefore, we will show in a third step, how addressing the principles of global inter- and intragenerational justice and the priority for the basic needs of the world's poorest sheds new light on the ethical reflection of AC. Thus, raising ethical issues within the Sustainable Development framework fosters the demand that AI technologies, much more than to date, must follow pathways that serve *all* humans and avoid discrimination and exclusion.

2 Overview on general ethical issues

Ethical issues that apply to all AI technologies (Hagendorff 2020) are also relevant in the context of AC. General points that are central for the specific ethical debate on AC as well are: 1) Protection against discrimination and equity of access and 2) protection of the user-privacy and cybersecurity. Some further points come up particularly when vulnerable people are involved, as in the case of care and education scenarios, with a focus on elderly people in the former and young people in the latter. In these cases, questions concerning 3) autonomy become especially relevant, as many elderly people as well as young children – and other people in certain contexts – are hardly able to take autonomous decisions . In addition, the characteristics of AC technology give rise to specific implications of general aspects, such as the problem of deception, the risk of stigmatization, and a more severe potential for misuse of the data.

2.1 Protection against discrimination and equity of access

Basically, when designing AC systems, it must always be kept in mind that, first, the underlying datasets are oftentimes biased and second, emotions are not value neutral. This point has already been discussed extensively in the literature. The datasets for training AI-systems mainly consist of white people, thus sometimes leading to problematic results, for instance, in search-engines (Makhortykh et al. 2021) or face-recognition-systems (Cavazos et al. 2021) when people of color interact with these systems. Even if developers are aware of that problem, it takes time and a lot of effort to enrich the existing datasets (Endrass et al. 2013) or develop new ones – time and effort that need to be financed and are part of a highly competitive time-critical economic field of innovation. Regarding AC, an additional short-coming needs to be kept in mind. Development and training of the algorithms are mainly based on the scale of universal emotions as suggested by Ekmann (1999), entailing six basic emotions: anger, surprise, disgust, enjoyment, fear, and sadness. However, in complex situations of social engagement, emotional settings are much more diverse than that.

The second aspect of possible discrimination and stigmatization poses a special challenge in the context of AC. The description or processing of emotions usually includes an evaluation of these emotions. Therefore, using the emotional data entails moral deliberating, which is inevitably done by the machine as well. This becomes highly problematic when people are categorized based on such an artificial procedure, like in the case of semi-intelligent information filters (SIFF systems). These are used to evaluate and interpret a set of data to draw conclusions about persons, marking them to behave suspiciously or be ready to act aggressively (Cowie 2015), conclusions that might have serious consequences.

Both mentioned sources of possible discrimination and stigmatization require intense attention and awareness, on the side of the developers and producers, as well as on the user side. The limitations of AC technologies must be understood and communicated thoroughly, which means that all members of society must be not only informed but educated accordingly.

Intense training and a broad corresponding ethical reflection would also serve to promote the development of applications that might be able to benefit everyone in society in a more suitable way. AC technologies can, for example, help to reduce discrimination by using culturally sensitive systems (Endrass et al. 2013). If developers were able to adapt artificial systems ("agents"[2]) more clearly to cultural settings in which they are to be used, awareness of – and maybe even more respect for – these differences would increase, presumably also on a global North-South scale.

This is an important aspect for AC technologies in *education* as well, where one of the main goals of the developers is to improve the usability of digital technology and the effectiveness of the learning experience. The (rudimentary) emotional sensitivity of the technology should enable people to use it more easily and in a more approachable way, so that users could take better advantage of the benefits the technology offers (Cowie 2015). This includes, for example, the possibility for the artificial AC system to react directly to frustrations on the part of the users, which is especially relevant in the educational context (Troussas 2020: chapter 5). Furthermore,

[2] We are preferring the rather neutral term artificial system for practical reasons and to avoid misunderstandings. In the literature on AI, often parlance is about artificial "agents" without discussing the conceptual implications. We do not consider these systems to possess agency according to a philosophical understanding of the term, linking actions to purposefulness and/or moral responsibility. If one day strong AI machines might be developed which are able to make decisions on their own, one would have to discuss further under which circumstances those machines could be (moral) agents. Yet, to date, AI technologies are not since they 'act' upon programmed algorithms. We are aware that this perspective is contested, and other scholars have different opinions on that (cf. Loh 2019: chapter 2.1).

emotionally sensitive technical assistants could help break down barriers to communication and access, enabling a more inclusionary approach towards diverse people. Emotion-sensitive language assistants could benefit people, for example, by enabling them to better cope with everyday life or by reducing communication barriers caused by different national or technical languages (Burchardt/Uszkoreit 2018). Access to various interfaces could be made easier for less technology-savvy people if these systems responded to negative reactions from users and at the same time presented a friendly and sympathetic counterpart (Cowie 2015). In addition, culturally specific adaptations could help to mitigate or overcome cultural hurdles in education scenarios as well. Furthermore, AC technologies are used in different training scenarios, for instance, for patients suffering from autism spectrum disorder (Obe et al. 2020) or people that face difficulties in stressful social settings (Schneeberger et al. 2019). Taken together, AC could contribute to increasing access to educational content within societies and on a global level. By this, AC could strengthen the potential for vulnerable groups to participate.

However, these advantages must be treated with caution, as AC systems could also have the exact opposite effect. This is especially the case when access to the technology is distributed unevenly, demands of transparency are not met or discriminating biases are not revealed. Furthermore, it must be critically questioned whether the enormous development effort can be justified in view of the mentioned ethical challenges, since it is doubtful how big the positive impact of AC systems will be. Therefore, expanded access for less privileged persons must be supported politically as well and goes far beyond the realm of technology development and distribution.

2.2 Protection of user privacy and cybersecurity

In the case of technical systems that record and analyze human emotions, a particular vulnerability is at stake for all users since the processed data is fundamentally intimate, sensitive information. In every-day situations within the public sphere, we show emotions and read the emotions of others frequently. However, in these situations we have some influence on what we want to show and what not, especially in cases we know we are recorded. In the case of AC systems, it is mandatory for the systems to work properly that we do not hide or fake out emotional states. And as is the case for all digitized data, these states are distributable and usable for other than the intended purposes. The demands and difficulties concerning data privacy have been discussed broadly in several contexts which already led to international standards[3]. In a justice-based approach, here individual privacy links up to social issues of tracking and treating the emotions of certain ‘suspec’ groups in an even more discriminatory way.

But, because of their intimacy, mass recording, generation, dissemination, and commodification of emotions through AC in the areas of digital imaging platforms and online transaction platforms are particularly critical and require special regulation (Stark/Hoey 2021), Here, it is plausible to follow the precautionary principle[4] (Andorno 2004) to rather slow down the speed of implementation of the AC technologies processing emotions in order to allow for broader technology and ethical assessments. In a practical sense, handling the information generated by these systems with the same caution as it is the case for medical data would allow for a high(er) ethical and legal status of protection.

For the use of digitized medical data, strict regulations already exist and could be adapted.

However, with increased networking and automatic processing, even these regulations face problematic limits. Particularly disputable is the topic of informed consent for (non- disclosure/processing, see Jörg 2018) as the information about all the processes involved is particularly complex and hard to understand even for healthy adult lay people. These problems also occur in the context of AC technologies, as has been already reported in the case of psychological treatment with AC support (Nicholas et al. 2020).

2.3 Autonomy

Factors seen as important for the autonomy of human persons is the ability of self-determination and being as independent in one’s decision making as

[3] The United Nations digital strategy (2022-25) lists guiding principles to digital transformation: https://digitalstrategy.undp.org/#Guiding-Principles, last access: 02.06.2022.

[4] The precautionary principle addresses situations where caution in the light of uncertainty should be given (Jordan/O’Riordan 2004). For example, the EU works with this definition: "The precautionary principle applies where scientific evidence is insufficient, inconclusive or uncertain and preliminary scientific evaluation indicates that there are reasonable grounds for concern that the potentially dangerous effects on the environment, human, animal or plant health may be inconsistent with the high level of protection chosen by the EU". (COM 2006: 90)

possible.[5] We take such forms of procedural independence as a requirement to be considered autonomous. In the case of AC machines, if third parties such as the operators of these machines or virtual systems, have information about or even access to emotional states of persons, this could have an influence on procedural independence (Baumann/Döring 2011), for example, by limiting the options for action. If I knew that a machine recognizes my emotional state, I might act differently than if this information were unknown. It might even be best to not consider some courses of action at all. As long as I am aware which different possibilities of action I have, this is not so much of a problem. However, people who are inexperienced in dealing with AC machines or cannot understand the technical limitations, e.g., children or individuals with mental disabilities or elderly people, are more vulnerable to unintended side-effects of their actions.

As a second point, such vulnerable groups might form unintended and/or unwanted bonds to artificial systems (Wilks 2010) in cases where the machines simulate emotions. To enter such an emotional relationship might limit the scope of self-determination and independent decisions as well.

Moments of misinterpretation can be challenging for autonomous action and decision. Those might occur when virtual systems show options for action that cannot be easily derived from the emotions they simulate at the same time. Or machines interpret emotions they recognize in an inadequate way. In such a case, it is hard for users to decide between different options for action because of not having decent information (Beavers/Slattery 2017). This problem is exacerbated if the machine relies on speech and text files from the Internet for the underpinning of options for action as the genesis and trustworthiness of such information cannot be recognized or verified instantly by the user.

Another form of possible deception can be described as *pars pro toto* acting, which is especially poignant with vulnerable groups. Imagine a teaching bot that gives the impression of caring. This might strengthen the bot in its

[5] We are using this rather basic understanding of the autonomy of human beings to illustrate some effects AC might have on our abilities of self-determination. In this context, the enduring and complex debate about different concepts of autonomy within philosophical ethics does not have to bother us for our limited purpose here. Furthermore, here, we are not discussing the question in what sense a machine can become an autonomous agent and what has to be the case to assign morality to robots or other artificial agents (cf. Loh 2019: chapter 2.1).

function. However, it might happen that users infer from one type of caring-behavior (e.g., reacting to the emotion of frustration while struggling with a math-problem) that the device is also capable of other caring contexts (e.g., dealing with the loss of a pet), which it might just not be (Cowie 2015).

Accordingly, it must be ensured for the field of *education* (as well as in the fields of gaming and care) that users have clear options for action. Furthermore, it needs to be transparent for which fields of action the artificial systems were developed and where their limits are, so that users can form realistic and binding expectations.[6] As in the case of all AI technologies, it is essential to assess the various possibilities of intentional or unintentional influence and manipulation, thus preventing corresponding misuse. If this fails, a loss of autonomy of the human actors, which goes hand in hand with possible manipulation, can be expected. In addition, the aspects of understandability and control of the technology, which is strongly emphasized in AI guidelines (Hagendorff 2020), must be taken seriously. The greatest possible transparency of the modes of operation should be given to ensure the most informed use possible. This also and especially applies to the target group of children, young adults, and mentally impaired patients due to their vulnerability. Therefore, the well-established discourse about informed consent, with all its intricate questions concerning vulnerable persons, must urgently be initiated for AC.

3 Anthropological perspectives

Until recently, emotions and emotionality have been regarded as a unique feature of humans or all living sentient beings (Manzeschke/Assadi 2019). At the same time, emotions are private in the sense that they cannot directly be accessed from the 'outside'. Artificial systems that can record as well as simulate human emotions more and more convincingly thus might challenge

[6] In this context it is discussed controversially whether it is helpful or not to design robots or avatars as humanlike in their appearance as possible. The Uncanny Valley Thesis (Mori et al. 2012) states that up from a certain point of realistic human appearance people tend to become afraid of artificial agents. However, as the Uncanny Valley is being questioned (Bartneck et al. 2007), it would go too far to elaborate on the topic more deeply in this chapter.

those basic anthropological assumptions.[7] This is relevant from a philosophical perspective, as several classical concepts of personal relations and capacities could be affected, like intersubjectivity, friendship and authenticity.[8] At the same time, it is also highly relevant from a societal perspective. All these concepts entail normative assumptions and moral obligations that not only affect singular people, but also societal ideas which in turn can affect societal transformations. How our understanding of what makes human life human, the *conditio humana* (Arendt 1998; Plessner 2003), can be affected by AC, we will elaborate on in the following.

Intersubjectivity, for instance, refers to the space of shared meanings between subjects that emerges through interpersonal exchange (Husserl 1973). These shared meanings can be understood as foundations of the social world, constituted by its members. So far, this social world has been shaped or created by human beings in dialogue (Buber 2009). If avatars or robots enter an (also) emotionally meaningful dialogue, then it is necessary to discuss whether our standard conception of intersubjectivity must be changed or extended. Furthermore, the question arises which consequences intersubjectively participating machines in our lifeworld would have on our understanding of the concept of a morally responsible person (Brand 2015). Ethically speaking, we must consider solidarity and inclusion as well as diversity-aspects regarding robots and other machines in both directions. This might seem far-fetched but the ongoing debate on demands for assigning a moral status to robots (Decker/Gutmann 2012; Loh 2019) and, accordingly, robot-rights (Gunkel 2021; Gellers 2021) poses some challenging insights. This brings us to the question if it would not be even more urgent to reflect and work on solidarity, inclusion, and diversity of *all* humans in society.[9]

Furthermore, the emotional and parasocial human-machine relationship promoted by AC might complement interpersonal relationships; a contested follow-up question is whether und under which circumstances this could also

[7] This is also connected to the discussion of what might happen if machines someday pass the Turing-Test.

[8] It might even affect our understanding of the capacity of making moral judgments. Emotions are discussed to take an important part in making moral decisions and having moral convictions. If artificial agents form emotionally based judgments, they might come closer to form moral judgments that are of a similar quality like human ones (Cowie 2015; Baumann/Döring 2011).

[9] We are fully aware on the debates on including sentient non-human beings into the moral community, but this is beyond the scope of this paper (cf. Bossert 2022).

lead to replacement of the latter. In this context, the concept of friendship, and along with it also the concept of care, might have to be reconsidered. Avatars or robots as so-called companions are becoming much more realistic both in the care sector and in the education sector as part of the development of AC. A Manzeschke and Assadi (2019: 170) point out that up to now people have been dependent on the functionality of machines of all kinds, insofar some form of dependence already exists. However, if people become emotionally dependent on artificial systems, then these machines might even set other standards in the sense of appropriateness of emotional response. This would open a shift or new dimension of the *conditio humana*. Another changing concept could be intimacy, if machines were able to collect, store, process and transmit human emotions indefinitely. This challenge already arises in the context of private conversations as well as sounds of everyday life that systems like Alexa and Siri permanently record.

Such developments again raise the question of solidarity, inclusion, and diversity. Do we really want to solve societal problems in the care and education sectors by developing simulations of human interaction? Would it be more suitable to work on human based solutions, especially when vulnerable people are concerned? Obviously, the concept of a good life (*eudaimonia*)[10] also must be addressed in this context as a rich social life and stable personal relationships are seen as a part of it, as among others, Turkle (2015) has pointed out with a critical view on the digital age.[11]

Another example of how AC might change some aspects not only of the c*onditio humana* but of our worldview is the human relation to a 'real' or 'natural' environment. This might change if AC-enhanced environments become more 'non-artificial' in the sense that the artificial systems seem more and more human to us by simulating emotions. The appeal of AC-enhanced environments could thus be enhanced and ultimately be preferred over non-AC-enhanced environments. For example, students who already have certain difficulties with human social contacts could experience them even more demanding, if they can rely on personalized avatars fulfilling exclusively their

[10] The discussion about what constitutes a good life is as old as philosophy and accordingly complex. We do not hold it necessary for the purpose of this chapter to go into details. However, we see Martha Nussbaum's Capability Approach as a promising framework to think about this question (Nussbaum 2000; 2007).

[11] We cannot go into the connected issues of objectophilia/object sexuality with regard to AI systems here.

wishes. With this, wishes are fulfilled without the necessity of – sometimes complex and demanding – social interaction. This can be seen as problematic if one holds that social competencies and at least some, even demanding, interactions with fellow humans are valuable and indispensable. The previously mentioned aspects, especially of inclusion and a good life are therefore affected here again.

These bundles of questions need to be further illuminated – not only but also – by in-depth philosophical research, for example, by analyzing the terms and concepts that might change as well as the anthropological implications and implications for moral psychology. What effects such changes might have cannot be foreseen in detail now. Nonetheless, we must reflect upon the ethical implications these developments might have. This should be done for the individual, the organizational, and the societal level (Manzeschke/Assadi 2019). The aspects we want to shed light on here – namely solidarity, inclusion, diversity, and further questions of good life –, are thereby located on the societal level. When addressing them, questions of governance must not only be discussed but also decided. Therefore, we need a normative framework for ethically reflecting AI developments that fundamentally include societal points of view. We suggest doing so by implementing the Sustainable Development perspective into the deliberation.

4 The Sustainable Development Perspective

One might ask why the Sustainable Development (SD) framework is a feasible choice regarding an (ethical) evaluation of AC technologies. Before answering this question, we shall highlight some important aspects of the SD concept.

Numerous academic as well as political documents on SD refer to the so-called Brundtland Report of the World Commission on Environment and Development (WCED 1987). This report brought the two political agendas of development and environmental conservation into a joint focus, two fields that had mostly been treated as contradictory to each other. Both are discussed as belonging together in the concept of SD, while setting the principle of inter- and intragenerational justice as the ethical foundation of SD. It reads:

> Sustainable Development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs. It contains within it two key concepts: the concept of

> 'needs', in particular the essential needs of the world's poor, to which overriding priority should be given; and the idea of limitations imposed by the state of technology and social organization on the environment's ability to meet present and future needs. (WCED 1987: chapter 2.1)

Inter- and intragenerational justice as a principle of equity is complemented by a priority principle regarding the basic needs of the world's poorest – and both justice aspects being explicitly framed by the limits of the non-human environment (ecological systems connected with the socio-technological systems). In the following, we will strengthen our argument – that a normative framework of SD is a suitable approach for AC technologies – by demonstrating how these principles can address the ethical challenges described in section 2. This is important because solidarity and inclusion, diversity, the influence on ideas of 'good life', the need for quality (digital) education, the planetary to local environmental conditions[12], and the question of governance all come to the table when deliberating if, why, and how machines should be trained to analyze and simulate human emotions at all.

4.1 Global inter- and intragenerational justice

Taking the principle of inter- and intragenerational justice seriously also means to strive for just societies today and in the future. We hold taking care of emotional needs of all members of societies as one important aspect for being able to achieve and sustain such just societies that foster a good life. Here, we cannot discuss in detail the – biologically, psychologically, as well as philosophically – complex term 'emotional needs'. Following the capability approach of Nussbaum (2000; 2007), we broadly understand emotional needs as, among others, the need to feel safe and being free from abuse, to be cared for, to be respected and to be self-effective.

For addressing these emotional needs, it is important to analyze from a societal perspective how a lack of emotional needs can be avoided or dimin-

[12] As every AI technology, AC requires large amounts of energy. The production of digital end devices needs resources, which often are rare and nonrenewable like rare earths, and the development of these technologies is responsible for an enormous amount of CO2 emissions (van Wynsberghe 2021). However, in this chapter, we do not focus on these aspects, which by no means should undermine their importance.

ished and what measures are best taken to cope with problems.[13] It might, as a general orientation, be more fruitful to invest in educational structures that are flexible enough to deal on an interpersonal level with emotional needs then to develop an emotion-simulation robot (the usefulness of some specific AC-applications and constellations notwithstanding).

Further, to take emotional needs seriously, the problem of possible user deception is to be considered in the development stage as well as in the implementation of AC technologies. This is necessary due to the danger of abusing an emotional need if the deception is not recognised as such. In order to prevent this from happening, transparency must be maintained. One way of avoiding unintentional possibilities of deception is to allow future users to participate in the development (Cowie 2015: 340).

When investing in technology development, a SD framework calls for focussing on the development of technologies which ease processes to strengthen public welfare. In line with this, it also calls for striving for *solidarity* with as well as the inclusion of *all* humans. Regarding AC technologies this means – as for all AI technologies – developing algorithms which do not revert to racist, sexist, anti-disabled, or other forms of discriminating biases. Instead, one needs to develop culturally sensitive systems (while avoiding discrimination and stigmatization; cf. section 2.1) and systems that mirror diversity (the appearance of the systems play an important role here, but also diversity in interaction should be fostered, as long as the diverse ways of interaction are evaluated as being helpful and valuable).

With this comes the problem that AC technology can reinforce discrimination and stigmatization (e.g., in the case of semi-intelligent information filter (SIFF) systems; cf. section 2.1). However, AC technologies can also serve to mitigate cultural differences, for example, in the area of intercultural training settings or through the use of appropriately sensitive learning companions. For being useful for all people, AC technologies appear to be worthy of support or expansion if they ensure greater access to digital or virtual systems,

[13] When discussing the question if and how AC technologies can foster or hinder a good life for as many individuals as possible, one also needs to investigate the question, if (in some cases even unrecognized) simulation of emotions or relations is detrimental to a good life or not (see above). We cannot provide a general answer to this question, especially since that hinges on the setup and context of the specific AI-system, yet we would point out the importance of critical perspectives, which Turkle (2015) and others have elaborated.

if the use of digital systems causes less stress or unpleasant experiences for people, or if cultural barriers can be made visible or be removed.

In sum, AC technologies should be built and used for enabling and empowering people – an aspect which a) needs to become an important litmus test for technologies to be evaluated ethically sound and useful and b) to meet the requirements given by a SD framework, namely, to comply with the principle of inter- and intragenerational justice. In such a way, the technology comes closer to enable good lives for (as many as possible) people.

On the other hand, to truly fulfill intragenerational (in the sense of global justice) means to seriously include the needs of the world's poorest. Without this, global justice cannot be realized. In relation to all technologies – including AC – this means developing them in a way that the technologies meet basic needs and enable basic social participation, rather than being adapted to luxury-oriented needs. In concrete terms, this means, for example, that AC technologies should be developed and used to enable accessible quality education in as many parts of the world as possible, rather than being used, for example, to improve micro-targeting for companies by assessing the emotions of potential customers. According to the Brundtland Report (WCED 1987), the needs of the world's poorest must be prioritized. This prioritization must be mirrored in technology development.

4.2 Education

Nearly 30 years after the publication of the Brundtland Report and following the spirit of the UN-Agenda 21 of the Summit for Environment and Development in Rio de Janeiro 1992, the United Nations' Sustainable Development Goals (SDGs) of 2015 have been agreed upon, which should be implemented in all states to transform societies into more sustainable ones.[14] As pointed out, we hold that education plays an important role for achieving intragenerational justice (cf. section 4.1) and we showed that AC technologies have a high potential to be used for education (cf. section 2.1). The SDGs explicitly address education in SDG No. 4: "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all". We want to stress that the full spectrum of education – which mostly is not considered in the

[14] For an interesting overview of the linkage of AI and the SDGs cf. Vinuesa et al. (2020) and Sætra (2021). Yet, both are not discussing AC technologies.

AI guidelines (Hagendorff 2020) – is of particular importance in the development of AC technologies. There is a quasi-double educational mandate for developers and users. Developers must be able to know, reflect and communicate the ethical implications of their systems. Users, on the other hand, must be able to inform themselves about the possibilities and limitations of the systems they use and to understand this information. Furthermore, developers must be able to recognize and reflect on ethical aspects of their work, especially regarding vulnerable groups, as it is often the case within the field of education. They need to be empowered to make their motivations and goals transparent to the public to contribute to an informed societal debate. This is especially true when it comes to particularly sensitive applications, such as dealing with emotions and vulnerable groups. All this is important and needs to be considered in relation to digital education programs, which have to be adjusted accordingly to serve establishing just societies.

Accordingly, these ethical and scientific communication skills should already be addressed during the training of further researchers and developers. If this were the case, then – at least in a roundabout way – more ethical reflection could also be incorporated into the development processes of commercial providers than has been the case to date.

4.3 Governance

The overall focus on justice highlights the whole set of principles such as equity of access and distribution, discrimination, and diversity, as well as education as central enablers for developing and using AC devices in a positive way. This eventually leads to governance questions in at least three main aspects.

First, with AC highly sensitive personal data can be collected, copied, and distributed. Therefore, data about emotions should be treated like medical data in general, granting the same high security and transparency standards to avoid misuse and respect the autonomy of the users. In this context, the fundamental question of the interaction between science and politics arises, as well as the question of social and political regulation. One can certainly ask whether, for instance, the further development of SIIF systems is socially or politically desired and whether AC should be permitted for free entrepreneurial use. And one must make sure that especially vulnerable persons are safe to use the technology. However, the decision to treat AC data as medical data is ultimately a governance decision.

Second, the discussion of AC technologies shows how urgently the dual educational mandate that arises in connection with AI technologies must be perceived. There is a need both to implement ethical competencies in the education of researchers and technicians, as well as to further expand digital education for all members of our society. This, again, entails a bundle of diverse governmental decisions, if taken seriously. Within the training of researchers and developers, for example, modules must be embedded that convey ethical competences from the start. When one remembers the effort it often takes to communicate digital knowledge in schools, it is obvious how crucial governmental decisions are in this area.

Third, the need to do justice to the world's poorest means that politics must not be oriented exclusively towards the interests of its own population. It must also consider the effects of local actions on other parts of the world. Measures should be pushed which not only do not have destabilizing consequences for the Global South – through, for instance, resource exploitation or selling of technologies to non-democratic regimes –, but which have the potential to benefit people in the poorest regions of the Global South. This requires highly complex governance activities in general and in the entire field of AI, including the field of AC (for regulatory options to minimize scenarios of AI damaging the SDGs cf. Truby (2020)). As always is the case in technology assessment and ethics, the danger is to look at AI-systems mainly from a technology-driven perspective. Taking into account SD as a guiding ethical groundwork, one would shift towards i) problem-driven and ii) cause-oriented approaches and ask how far AC-systems really could help here (Erdmann et al. 2022). This would also have implications for the way governance of such technologies should be organized and structured in the first place, namely not by separate but integrative regulatory works.

5 Conclusion

The normative framework of Sustainable Development as a basis to investigate and evaluate Affective Computing shows that precisely this perspective can – and should – complement, if not integrate, the common AI-ethical considerations, since it meaningfully broadens and at the same strengthens ethical considerations.

With the strong focus on the principles of inter- and intragenerational justice underlying the SD framework, the recognition and simulation of emotions by machines should be done in a way that empowers all humans and reduces the risk of deception to the lowest possible level.

Moreover, if the prioritization of the basic needs of the world's poorest is taken seriously, this would prevent AC technologies from being used primarily to satisfy market-oriented interests of the Global North, like for example the development of even more realistic computer games or targeted advertising that not only addresses an individual's interests, but also their current emotional state. AC must then be used in a way that promotes fundamental interests such as equal participation in society. This can be implemented for example by using AC for quality digital education programs. However, the use of AC in an ethically acceptable or even desirable way does not only serve people in the Global South. It will also be useful to form more just societies in the Global North if, for example, data about emotions is treated like sensitive medical data in the development and use of this technology. While this is necessary for the general usage of AC, it is specifically important in the field of education. Here, AC has a high potential, but this field also usually affects vulnerable groups. It will also benefit *all* societies if programming takes into account cultural specificities and the avoidance of discrimination and stigmatisation, and if emotion recognition and simulation by machines is developed to provide digital education and training scenarios at a level that would be more difficult without this technology.

References

Andorno, Roberto (2004): "The Precautionary Principle: A New Legal Standard for a Technological Age." In: Journal of International Biotechnology Law 1/1, pp. 11-19.

Arendt, Hannah (1998 [1958]): The Human Condition. 2nd edition, Chicago: University of Chicago.

Bartneck, Christoph/Kanda, Takayuki/Ishiguro, Hiroshi/Hagita, Norihiro (2007): "Is the Uncanny Valley an Uncanny Cliff?" In: RO-MAN 2007 – The 16th IEEE International Symposium on Robot and Human Interactive Communication, pp. 368-373.

Baumann, Holger/Döring, Sabine (2011): "Emotion-Oriented Systems and the Autonomy of Persons." In: Paolo Petta/Catherine Pelachaud/Roddy Cowie (eds.), Emotion-Oriented Systems: The Humaine Handbook, Heidelberg: Springer, pp. 735-752.

Beavers, Anthony F./Slattery, Justin P. (2017): "On the Moral Implications and Restrictions Surrounding Affective Computing." In: Myounghoon Jeon (ed.), Emotions and Affect in Human Factors and Human-Computer Interaction, London: Elsevier Academic, pp. 143-161.

Bossert, Leonie N. (2022): Gemeinsame Zukunft für Mensch und Tier – Tiere in der Nachhaltigen Entwicklung, Baden-Baden: Karl Alber.

Brand, Cordula (2015): "Wie Du mir so ich Dir: Moralische Anerkennung als intersubjektiver Prozess." In: Robert Ranisch/Sebastian Schuol/Marcus Rockoff (eds.), Selbstgestaltung des Menschen durch Biotechniken, Tübingen: Narr Francke Attempto, pp. 21-33.

Buber, Martin (2009): Das dialogische Prinzip, Gütersloh: Gütersloher.

Burchardt, Aljoscha/Uszkoreit, Hans, eds. (2018): IT für soziale Inklusion: Digitalisierung – Künstliche Intelligenz – Zukunft für alle, München and Wien: De Gruyter Oldenbourg.

Cavazos, Jacqueline G./Phillips, P. Jonathon/Castillo, Carlos D./O'Toole, Alice J. (2021): "Accuracy Comparison Across Face Recognition Algorithms: Where Are We on Measuring Race Bias?" In: IEEE Transactions on Biometrics, Behavior, and Identity Science 3/1, pp. 101-111.

COM (2006 [1995]): "Communication from the Commission of 2 February 2000 on the Precautionary Principle (COM (2000) 12.02.2000, P. 1)." In: Philippe Sands/Paolo Galizzi (eds.), Documents in European Community Environmental Law, 2nd edition, Cambridge: Cambridge University, pp. 90-115.

Cowie, Roddy (2015): "Ethical Issues in Affective Computing." In: Rafael A. Calvo/Sidney D'Mello/Jonathan Gratch/Arvid Kappas (eds.), The Oxford Handbook of Affective Computing, Oxford library of psychology, Oxford: Oxford University, pp. 334-348.

Decker, Michael/Gutmann, Mathias, eds. (2012): Robo- and Informationethics: Some Fundamentals, Wien, Berlin and Münster: Lit.

Devillers, Laurence (2021): "Human-Robot Interactions and Affective Computing: The Ethical Implications." In: Joachim von Braun/Margaret S. Archer/Gregory M. Reichberg/Marcelo Sánchez Sorondo (eds.), Robotics, AI, and Humanity: Science, Ethics, and Policy, Cham: Springer International, pp. 205-211.

Ekmann, Paul (1999): "Basic Emotions." In: Tim Dalgleish/Michael J. Power (eds.), Handbook of Cognition and Emotion, Chichester: Wiley, pp. 45-60.

Endrass, Birgit/André, Elisabeth/Rehm, Matthias/Nakano, Yukiko (2013): "Investigating Culture-Related Aspects of Behavior for Virtual Characters." In: Autonomous Agents and Multi-Agent Systems 27/2, pp. 277-304.

Erdmann, Lorenz/Cuhls, Kerstin/Warnke, Philine/Potthast, Thomas/Bossert, Leonie/Brand, Cordula/Saghri, Stefani (2022): Digitalisierung und Gemeinwohl – Transformationsnarrative zwischen planetaren Grenzen und Künstlicher Intelligenz, Texte 29/2022, Dessau-Roßlau: Umweltbundesamt.

Gellers, Joshua C. (2021): Rights for Robots: Artificial Intelligence, Animal and Environmental Law, Oxon and New York: Routledge.

Gunkel, Harald (2021): "Robot Rights – Thinking the Unthinkable." In: John-Stewart Gordon (ed.), Smart Technologies and Fundamental Rights, Leiden and Boston: Brill Rodopi, pp. 48-72.

Hagendorff, Thilo (2020): "The Ethics of AI Ethics: An Evaluation of Guidelines." In: Minds and Machines 30/1, pp. 99-120.

Husserl, Edmund (1973): Zur Phänomenologie der Intersubjektivität: Texte aus dem Nachlass. Husserliana: Edmund Husserl. Gesammelte Werke Vol. 13, 14, 15, Den Haag: Martinus Nijhoff.

Jörg, Johannes (2018): Digitalisierung in der Medizin: Wie Gesundheits-Apps, Telemedizin, künstliche Intelligenz und Robotik das Gesundheitswesen revolutionieren, Berlin and Heidelberg: Springer.

Jordan, Andy/O'Riordan, Tim (2004): "The precautionary principle: A legal and policy history." In: Marco Martuzzi/Joel A. Tickner (eds.), The Precautionary Principle: protecting public health, the environment and the future of our children, Copenhagen: World Health Organization, pp. 31-48.

Loh, Janina (2019): Roboterethik: Eine Einführung, Berlin: Suhrkamp.

Makhortykh, Mykola/Urman, Aleksandra/Ulloa, Roberto (2021): "Detecting Race and Gender Bias in Visual Representation of AI on Web Search Engines." In: Ludovico Boratto/Stefano Faralli/Mirko Marras/Giovanni Stilo (eds.), Advances in Bias and Fairness in Information Retrieval. Communications in Computer and Information Science Vol. 1418, Cham: Springer International, pp. 36-50.

Manzeschke, Arne/Assadi, Galia (2019): "Emotionen in der Mensch-Maschine Interaktion." In: Kevin Liggieri/Oliver Müller (eds.), Mensch-Maschine-Interaktion: Handbuch zu Geschichte – Kultur – Ethik, Berlin: J.B. Metzler, pp. 165-171.

Mori, Masahiro/MacDorman, Karl/Kageki, Norri (2012): "The Uncanny Valley [From the Field]." In: IEEE Robotics & Automation Magazine 19/2, pp. 98-100.

Nicholas, Jennifer/Onie, Sandersan/Larsen, Mark E. (2020): "Ethics and Privacy in Social Media Research for Mental Health." In: Current psychiatry reports 22/12, 84.

Nussbaum, Martha Craven (2000): Women and Human Development: The Capabilities Approach, Cambridge: Cambridge University.

Nussbaum, Martha Craven (2007): Frontiers of Justice: Disability, Nationality, Species Membership, Cambridge and London: The Belknap Press of Harvard University.

Obe, Olumide/Akinloye, Folasade Oluwayemisi/Boyinbode, Olutayo (2020): "An Affective-Based E-Healthcare System Framework." In: International Journal of Computer Trends and Technology 68/4, pp. 216-222.

Picard, Rosalind W. (2000): Affective Computing, Cambridge and London: The MIT.

Plessner, Helmuth (2003 [1986]): Conditio humana. 4. Auflage. Gesammelte Schriften in zehn Bänden Vol. VIII, Frankfurt am Main: Suhrkamp.

Sætra, Henrik (2021): "AI in Context and the Sustainable Development Goals: Factoring in the Unsustainability of the Sociotechnical System." In: Sustainability 13/4, 1738.

Schneeberger, Tanja/Gebhard, Patrick/Baur, Tobias/André, Elisabeth (2019): "PARLEY: A Transparent Virtual Social Agent Training Interface." In: IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion, New York: Association for Computing Machinery, pp. 35-36.

Stark, Luke/Hoey, Jesse (2021): "The Ethics of Emotion in Artificial Intelligence Systems." In: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, pp. 782-793.

Troussas, Christos (2020): Advances in Social Networking-Based Learning: Machine Learning-Based User Modelling and Sentiment Analysis, Cham: Springer International.

Truby, Jon (2020): "Governing Artificial Intelligence to Benefit the UN Sustainable Development Goals." In: Sustainable Development 28/4, pp. 946-959.

Turkle, Sherry (2015): Reclaiming Conversation: The Power of Talk in a Digital Age, New York: Penguin.

van Wynsberghe, Aimee (2021): “Sustainable AI: AI for Sustainability and the Sustainability of AI.” In: AI Ethics 1/3, pp. 213-218.

Vinuesa, Ricardo/Azizpour, Hossein/Leite, Iolanda/Balaam, Madeline/Dignum, Virginia/Domisch, Sami/Felländer, Anna/Langhans, Simone Daniela/Tegmark, Max/Fuso-Nerini, Francesco (2020): “The Role of Artificial Intelligence in Achieving the Sustainable Development Goals.” In: Nature Communications 11/1, 233.

Wilks, Yorick (2010) “Introducing Artificial Companions.” In: Yorick Wilks (ed.), Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues Vol. 8, Amsterdam: Benjamins, pp. 11-20.

World Commission on Environment and Development (WCED) (1987): Our Common Future, Oxford: Oxford University.

Part III – Care, Love, Sex

Granny and the Sexbots
An Ethical Appraisal of the Use of Sexbots in Residential Care Institutions for Elderly People

Karen Lancaster

> "One resident, due to physical disability, is unable to masturbate himself. He confesses that he does become very sexually frustrated and I am at a loss as to how to address the issue for him. Unfortunately he is not in a position to be able to go and meet women of his own accord. I understand his needs, but do not know what I can legally do to assist him in his quest for sexual gratification." Quote from an eldercare nurse (Royal College of Nursing 2011: 4).

1 Introduction

By 2050, it is expected that over a fifth of the world's population will be aged over 60 (WHO 2018); this includes an expected 3.7 million people over the age of 100 (Stepler 2016). Although people's lives are increasing in duration, this does not necessarily mean there will be a corresponding increase in the number of years spent living independently. Care institutions for elderly people will be increasingly necessary, and people will spend a greater proportion of their lives in institutional care than ever before.

Noel and Amanda Sharkey (2012) wrote a paper entitled *Granny and the Robots*, in which they raise some ethical concerns associated with robotic carers for elderly people. The concerns they highlight include a reduction in human contact; feelings of objectification and loss of control; loss of privacy; loss of liberty; deception and infantilisation; and difficulties controlling the robots (Sharkey/Sharkey 2012). Sharkey and Sharkey lay out some preliminary groundwork for future ethical discussions of carebots for elderly people, and provide a brief cost-benefit analysis of the introduction of such robots. My project in this paper (and my title) echoes theirs, but in the sphere of sexbots for elderly people. Some of the issues identified by Sharkey and Sharkey are also relevant to sexbot usage. These include a reduction in human contact, objectification, and difficulties controlling the robots; these questions of social justice will be discussed forthwith in addition to the question of whether sex is a good to which people are entitled, and whether care institutions have a duty to cater for sexual needs.

Sex among elderly people is one aspect of life which remains taboo. I argue that the sexual appetites of elderly residents in care institutions ought to be catered for – not least because the residents themselves feel it is an important aspect of their care (Royal College of Nursing 2011: 11). Sex and intimacy also help to improve residents' mood, health, and general quality of life (Hajjar/Kamel 2003). Importantly, I suggest that sexbots are apt to provide this service (with some caveats and modifications).[1] First, I demonstrate that sexual feelings and behaviours are an important part of elderly people's lives. Following that, I provide a brief cost-benefit analysis of other sexual options for elderly people, noting that some options are potentially morally problematic or simply impractical. I argue that although sexbots will not be a panacea – indeed, they will create some ethical issues which need to be explored philosophically, and they will require robust safeguarding protocols – they have the potential to enhance the lives of residents in eldercare institutions, and their use can be implemented whilst minimising problems.

1.1 Assumptions and terminology

I use the term 'eldercare institution' (or simply 'care institution') to refer to some sort of private or state-run residential caring environment for elderly people; this is taken to be a structured, non-familial setting, where multiple elderly people are cared for. I refer to the people who live in the eldercare institutions as 'residents', although the term 'patients' may also be apt. The residents I refer to are taken to be in need of physical care due to frailty, disability, illness, infirmity, and other physical impairments, but to have full mental capacity (they are not suffering from dementia, learning disabilities, or severe

[1] Note that this paper is an ethical argument rather than a technical manual for roboticists. I do not show *how* sexbots for elderly people should be developed, but I do make some suggestions of modifications which would help to facilitate the use of sexbots by elderly people.

mental illness). Some cognitive decline is commonplace as people age, and so additional safeguards and philosophical exploration would be needed to protect people without full mental capacity, should they wish to use sexbots. I do not consider arguments pertaining to such people herein. I shall use the term ‘nurses’ to refer to care workers, nurses, nursing assistants, personal care aides, medical assistants, patient welfare assistants, and other such similar jobs which involve looking after elderly residents in care institutions.

My argument is intended to demonstrate that eldercare institutions (or their funding bodies) should provide the sexbots, rather than merely suggesting that residents should be allowed to bring their own sexbot into an eldercare institution with them. There would seem to be less to prove in such a case where a resident already possesses, uses, and cleans their own sexbot – although my argument herein can also support the suggestion that elderly people should be allowed to have their own sexbots too. Generally, the standard of necessity and safety is set at a higher level for equipment or services which are provided by an institution, compared to things which one provides for oneself. There may be separate arguments to be made in favour of allowing residents to use their own sexbots; the argument I make here is that eldercare institutions could and should provide sexbots for use by residents.

2 Sexual desire and activity in elderly people

Residents in eldercare institutions have a range of needs, and this results in them requiring care in a range of ways. Perhaps most obviously, elderly people have physical needs which must be met via physical care from nurses. These activities may include, but are not limited to, help with washing, dressing, toileting, moving about, using equipment (such as stairlifts), sitting and standing, eating, sleeping, and personal grooming. Elderly people – like everyone else – also have social and emotional needs which can be catered for via activities such as entertainment, games, leisure, music, day trips, learning new skills, socialising, and sports which are appropriate to their level of physical ability. It seems clear that an eldercare institution which merely looks after the physical needs of its residents (whilst ignoring their social and emotional needs) is failing its residents. Depression, anxiety, and loneliness are unfortunately fairly widespread among elderly people; it is estimated that around 40 per cent of eldercare institution residents experience depression, and many of these people will experience low mood *throughout* their time in institutional care, which is generally a number of years, until death (Social Care Institute for Excellence 2006, British Geriatrics Society 2018). Clearly, there is still much to be done to improve the social and emotional welfare of some of society's most vulnerable citizens.

Someone might be forgiven for thinking that the problem I outline herein is not really a problem at all: that elderly people do not have sexual desires, and having access to a sexbot would not improve their quality of life in any discernible way. However, there are certainly residents – like the one who was referred to by the nurse at the outset – for whom sexual frustration is a daily plight which makes life less pleasant than it needs to be. I believe that sexbots, properly utilised, could significantly improve the lives of such people. To the best of my knowledge, there are no eldercare institutions which provide sexbots for residents, so empirical evidence which demonstrates how great an improvement sexbots can make to the lives of elderly people is in short supply. What I do show in this paper is that sexuality is an important feature of physical and emotional wellbeing, and that sexbots are a viable outlet for sexual behaviour.

Although positive steps forward are being made to cater for elderly residents' social and emotional needs, sexual desire is often considered non-existent, unimportant, or even problematic. This sort of mindset is highly concerning, because an institution which ignores or trivialises sexual desire and sexual activity among its residents can cause deep unhappiness for the elderly people involved (Royal College of Nursing 2011: 7). Unfortunately, this ignoring, trivialisation, and even prevention of sexual behaviour among elderly people often does seem to occur in eldercare institutions.

There exists a pervasive and inaccurate view of an asexual old age, yet sexual desire and sexual satisfaction remain important for many people throughout their later years, including times when they are in eldercare institutions (Gott/Hinchcliff 2003, Hajjar/Kamel 2003, Franowski/Clark 2009). It is often unacknowledged that people over the age of 70 still have sexual desires, and can and do enjoy sexual activity. There are noticeable gaps in policy documents which could address sex among elderly people directly. For example, the Strategic Action Plan for Sexual Health (Public Health England 2015) makes little mention of sexual health in old age, instead focusing primarily on young people. Training bodies could educate nurses about the issue; however, it is commonplace for no training to be provided for nurses which could enable them to adequately deal with sexual desire among the elderly people in their care (Royal College of Nursing 2011: 3).[2] It would be possible for eldercare institutions to develop their own policies or formal guidelines on sexual activity among residents – however most institutions have nothing in writing, and just deal with sexual behaviour on an ad hoc basis (Tarzia et al. 2012: 610). More troublingly, studies strongly indicate that frequently, eldercare institutions do not have facilities or attitudes which support or allow for sexual expression among their residents (Roach 2004; Franowski/Clark 2009; Care Quality Commission 2020: 23). For example, eldercare institutions may not have lockable bedroom doors; bedrooms may only have single beds; or residents may not be permitted to sleep in other people's bedrooms – these rules can even apply to married couples living within the same institution (Royal College of Nursing 2011: 4). This means that even if a resident wants to have a sexual relationship with someone else, it would be very difficult or impossible to have private time or to spend the night together. Elderly people engaging in sexual activity with one another risk being disturbed by nurses, who may view the activity as an 'incident' which needs reporting or tackling, and nurses may even intervene to break up consensual sexual activity between two adults (Care Quality Commission 2020: 17).

The fact that sexual desire in old age remains a taboo subject (Care Quality Commission 2020: 27-33; Royal College of Nursing 2011: 3) does not demonstrate that such desire is non-existent, however. Although government strategies and institutional policies may not place much importance on the sexual desires and activity of elderly people, elderly people themselves feel that their sex lives *are* important. A study by Gott and Hinchcliff (2003) found that over 60% of respondents aged over 70 rated sex as 'moderately', 'very' or 'extremely' important to them. They also found that among the 40% who said it was not important or only a little important, a common reason cited for this lack of importance was the belief that they would not have sex again in their lifetime, whether due to ill health, disability or widowhood. It is possible that these respondents were trying to downplay the importance of sex in their lives, believing that it was no longer an option even if they wanted it. Most elderly people in institutional care believe that opportunities for sexual expression should form part of their care (Royal College of Nursing 2011: 11). This alone

[2] It is interesting – and disheartening – to note that this document on sex and sexuality in care homes has not been updated since its creation 11 years ago; other updated documents on care homes and nursing in the UK continue to make no mention of sex among older people.

should be sufficient cause for eldercare institutions to reconsider their policies and attitudes towards sexual activity among residents. Sex and intimacy are important features of elderly people's health and wellbeing; they help to improve mood, health outcomes, and quality of life (Hajjar/Kamel 2003). In short, sexual desire and sexual activity remain important for both mental and physical wellbeing while people are in eldercare institutions, just as they are important for younger adults living independently.

The lack of infrastructure, facilities, or policy guidelines is certainly problematic, but this need not be the case. There have been moves in recent years which demonstrate that sexual needs among elderly residents are being acknowledged and accepted more often by nurses. For example, the Royal College of Nursing recommends that residents should be permitted to explore sexual relationships with others (just as people are free to do when living independently), and that private spaces should be provided to enable residents to engage in consensual sexual relations with one another (or alone) if they so wish (Royal College of Nursing 2011: 3-6). Nonetheless, even if nurses and policy documents recognise that sexual desire exists among elderly people, many residents will still have sexual desires which go unmet.

3 Ways in which sexual desires can be met: Cost-benefit analysis

In this section I consider four possible non-sexbot ways in which elderly people could engage in sexual activity within care institutions, and I provide a brief cost-benefit analysis of each. None of these solutions – nor sexbots – is perfect for everyone, however some of these possibilities may be suitable for some residents. My argument in favour of sexbots should not be taken to suggest that these other options should be closed off to residents: the ideal situation would be one where elderly people in institutional care have a variety of ways in which they can explore their sexuality. I am simply arguing that sexbots could help elderly people for whom these possibilities are problematic, or who would simply prefer to have sex with a sexbot. The possibilities I consider are:

- Other residents
- Nurses
- Sex workers
- Sexual aids

I have noted above that few eldercare institutions have an environment which facilitates and allows for sexual activity (such as private rooms with double beds and lockable doors); these sorts of facilities would be required for several of these other possibilities, but in the interests of brevity, I do not repeatedly cite lack of privacy as a reason against these possibilities.

3.1 Other residents

Perhaps one of the best ways an institution can cater for sexual activity among its residents is by encouraging and facilitating intimate relationships among eldercare institution residents. Meeting an intimate partner in an eldercare institution brings with it the possibility not just of sexual release, but of love and companionship – something which is exceptionally valuable for emotional wellbeing (Hajjar/Kamel 2003). Roboticists are working towards producing robots which offer love and companionship for users (Anctil/Dubé 2019) – or at least the appearance of it. However, at present, a sexbot cannot provide anything like the same level of emotional companionship that a human being can, and for this reason, a human-human sexual relationship would seem to be an ideal option for residents who can manage it.

However, relationships within eldercare institutions are not a viable possibility for everyone. One reason for this is that there are far more women in eldercare institutions than there are men: in total, there are around three women for every man in institutional care in the UK – this rises to over four women for every man among residents aged 85 and over (Office for National Statistics 2014). Most people are heterosexual, and this gender imbalance greatly reduces the chance for heterosexual women to meet a suitable partner within an eldercare institution (options for homosexual people are also limited; heterosexual men are the only ones who have plenty of potential partners available). Even for those people lucky enough to meet a partner with whom they want to be sexually intimate, there still remain significant difficulties with actually *engaging* in sexual activity – after all, if residents were fully healthy, mobile and dextrous, they would not require institutional care in the first place. Elderly residents may well suffer from frailty, stiff joints, pain, sickness, and inability to physically exert themselves too much – issues which make sexual intercourse difficult or impossible. Chronic conditions such as these do not affect sexual desire (Kalra et al. 2011), and so even people who are suffering severe physical decline may still wish to – but be unable to – engage in sexual activity with others.

The lack of potential partners, and physical inability to engage in sex means that sex with other residents is simply not an option for most people in eldercare institutions.

3.2 Nurses

A different – but deeply flawed – possibility for sexual release is that nurses could provide sexual services for residents. Given that nurses often already engage with residents in intimate ways – for example, by dressing them, bathing them, and helping with their toileting needs – it may seem like only a small additional step for them to engage in sexual activity too. However, such a move would be fraught with problems – not least because it substantially changes the nature of the relationship between the carer and the patient. There would undoubtedly be many nurses who would view such a change to their job as a deal-breaker: if sex were to become a feature of the job, there would likely be an exodus of nurses leaving the profession, exacerbating the existing shortage of nurses.

Such a great change to the nature of nursing could also cause untold harms to the nurses themselves, such as stress, anxiety, and depression. If providing sexual services were to become a duty of the nurse, this effectively removes her right to sexual self-determination; she waives her claim to bodily integrity – something we generally view as important (Vandervort 1987). Many nurses would be unable to change their jobs, meaning that they would be stuck in employment effectively as a sex worker – a type of employment which is highly stigmatised, and has a range ethical and social problems associated with it (as we shall see in the next sub-section).

Furthermore, permitting, encouraging, or mandating sexual activity between nurses and residents would throw open the opportunity for abuse – something which would not be in the best interests of residents (or nurses). Elderly people are some of the most vulnerable members of society. Unlike children who will one day become independent and can speak out about their abusers, people in eldercare institutions will generally remain in the institution – or another like it – until death, meaning that they may never get the opportunity to report their abuser. Although statistics suggest that sexual abuse in institutional care is not commonplace (it accounts for just 3% of all abuse complaints; Care Quality Commission 2020: 6), it is probably underreported. Sexual behaviour between nurses and residents could seem to endorse or normalise sexual abuse, giving abusers the opportunity to claim that their sexual activity with a resident was consensual. Given the severe negative effects of sexual abuse in the eldercare sector, anything which could help abusers get away with their crimes is best avoided (National Center on Elder Abuse 2018: 2; Care Quality Commission 2020: 28-31; Age UK 2020, 2021). For these reasons, promoting sexual activity with nurses would seem to cause far more problems than it would solve, and is thus not a viable outlet for sexual activity.

3.3 Sex workers

A different option, then, could be the use of sex workers. Many sex workers are highly experienced and could provide a good standard of sexual pleasure for residents. Moreover, they would be able to work around disabilities or infirmities which residents may suffer from. There are enough sex workers available so as to provide some level of choice to residents – for example, there are men and women of different ages, ethnic groups, and sexual orientations who are sex workers.

The potential spread of sexually transmitted infections (STIs) is an immediate concern when considering using sex workers. Although this is an important consideration, this could be addressed via hygiene requirements, frequent STI testing, and the use of barriers such as condoms. However, there are other problems associated with the use of sex workers which would be more difficult to address.

Most crucially, institutions may rightfully be concerned about the pitfalls of endorsing prostitution. There is a huge debate about the morality of prostitution; it has been variously criticised as harmful, dangerous, objectifying, running contrary to Kantian notions of treating people as a mere means, reinforcing gender stereotypes, and legitimising sexual violence (Chevarie-Cossette 2017, Richardson 2016, Varden 2006, Westin 2014, Vicente 2016, Spector 2006, Thomsen 2015, Settegast 2018). Within the UK, for example, soliciting or paying for sex with someone who has been coerced into prostitution is a summary offence (even if one does not *know* that the sex worker was coerced into it) – although the police have the discretion not to prosecute such offenders (Crown Prosecution Service 2019). Prostitution is also often associated with human trafficking, violence, and drug and alcohol abuse – but even in the absence of these additional harms, endorsing prostitution remains morally questionable. Due to these reasons, prostitution is not something which any reputable eldercare institution – not least one which is funded by the state – could permissibly be involved with.[3] Aside from legal worries, there is often a very strong social stigma associated with using sex workers: it is often seen as undignified, lewd, or desperate. Residents may therefore be resistant to using a sex worker on ethical grounds (such as concerns about the welfare of the workers) or because they simply do not want to suffer the social stigma which can come with using sex workers.[4]

In spite of the problems associated with using sex workers, there have been cases in the UK of (usually young) disabled people using sex workers – encounters which were set up or facilitated by nurses (Stretch 2013; Ismail 2013; Little 2007). Organisations exist which facilitate meetings between approved sex workers and disabled people (TLC Trust 2021). Depending on how one regards the moral, social, and legal issues, the use of sex workers for elderly people could be a solution for some people.

3.4 Masturbation and sexual aids

It may be the case that some elderly people are able to masturbate successfully, while others may be unable to easily or successfully achieve any satisfaction from masturbation, but have sexual desires nonetheless. The male patient referred to at the beginning of this paper, who is unable to masturbate and feels very sexually frustrated (Royal College of Nursing 2011: 4), is one such example. With nearly 300,000 elderly people in institutional care in the UK alone (Office for National Statistics 2014), there are undoubtedly many other elderly people who feel the same sexual frustrations but have not spoken to a nurse about it and had their case recounted in a policy document. For some people who are unable to masturbate but prefer to be alone for their sexual gratification, a sexual aid of some sort could be a good idea.

[3] In the next paragraph I note some recent cases where disabled young adults *were* helped by their carers to seek out and use prostitutes. Of course, simply because it has taken place does not prove that doing so is permissible or unproblematic.

[4] There may also be a stigma associated with using sexbots which echo those associated with using prostitutes: that it is undignified, lewd, and desperate. This may change over time as robots become more commonplace, or it may not. Where robots have the edge over sex workers, however, is that we do not need to concern ourselves with the welfare of the robot (so long as it is not sentient), nor whether it was coerced, trafficked, drugged, blackmailed, or threatened into prostitution. There may nonetheless remain concerns over whether sexbots reinforce misogyny and objectification of women, as prostitution does (see Richardson 2016).

There are many types of sexual aids commercially available – battery-operated mechanical devices such as vibrators; silicon devices which mimic human genitals; swings to facilitate more comfortable movement; and a plethora of other creations to suit every taste. Some of these are capable of providing or helping to provide a satisfying sexual experience which could allow elderly people to explore their sexuality alone or with someone else. One benefit of such devices is that they are often cheap enough that residents could have one device (or multiple devices) to themselves, preventing the spread of sexually transmitted infections. Moreover, because many such sexual aids are small, they could be used discreetly and put away without nurses, other residents, or family members knowing of their existence; this would help to limit embarrassment of all parties, given that sex and masturbation are still taboo subjects among elderly people (Care Quality Commission 2020: 27-33; Royal College of Nursing 2011: 3).

One problem with such aids is that they may not always provide a wholly satisfying intimate experience, since they are generally an aid to masturbation rather than a substitute for sex. There may be large numbers of people who are unable to use sexual aids effectively, or simply do not wish to. Clearly, masturbation is not an option for all elderly people within institutional care – as Di Nucci notes when writing about sexbot usage by disabled people: "if masturbation were the solution to the problem, then we wouldn't have had a problem in the first place" (Di Nucci 2017: 77). Nonetheless, masturbation (with or without sexual aids) is something which at least some elderly people will be capable of engaging in while in eldercare institutions, and it would certainly be ideal if nurses could endorse and facilitate that where possible (for example, by knocking and waiting before entering a resident's room).

4 Arguments against sexbot provision

I shall now turn my attention to addressing some potential arguments *against* sexbot provision in eldercare institutions, and I shall give responses and solutions to these arguments as appropriate. First I deal with a general argument against sexbots (that sexbots are morally problematic); an argument which may be relevant to the use of sexbots in any part of society, not merely in eldercare institutions. The other arguments against sexbot provision are relevant specifically to care institutions for the elderly.

One ostensible reason why people may think sexbots have no place in eldercare institutions is that elderly people do not have sexual interests – however, since I have addressed this earlier and established that such a claim is simply false, I shall not address it again here: elderly people *do* have sexual appetites. There are also numerous pragmatic reasons why sexbot provision is questionable, such as cost[5], inconvenience, and sexually transmitted infections[6]. Some people might suggest that having sexbots in eldercare institutions would be inconvenient for nurses (maintaining and cleaning the sexbots would give them another responsibility, when they are already overworked). However, it seems only humane to suggest that the quality of life for elderly residents is important and should take priority over inconveniencing nurses (Sharkey/Sharkey 2012: 37). Sexbots might offend some people who consider them shocking and lewd, but this alone is not sufficient reason why they should not be supplied in eldercare institutions. If it is the case that sexbots can improve the quality of life for elderly people, and that their quality of life is important, then the idea of sexbot provision in eldercare institutions should be taken seriously.

Below, I consider and respond to several ways in which people might argue that sexbots should not be provided in eldercare institutions. These are: sexbots are morally problematic; elderly people would not want sexbots; we do not expect sexbots to be provided in eldercare institutions; people outside of institutional care do not have constant access to sex; sex is a want rather than a need; and not all desires can or should be catered for in institutional care.

Following these considerations, I address the question of whether sexbots are up to the task – that is to say, whether they can safely and satisfyingly provide a sexual outlet for elderly people, given the fact that sexbots are heavy, rigid, and passive recipients of sex. This is not strictly an argument against sexbot provision; rather, it is a caveat regarding the feasibility of sexbot usage by elderly people.

[5] The cost of a typical AI sexbot is currently around £2000-4000 (Shenzhen All Intelligent Technology Co. Ltd 2022, Smart Doll World 2022). Inanimate sex dolls come in much cheaper, at around £500 (Realdoll 2021), but these may be less useful to elderly people with limited mobility.

[6] With proper cleaning, sexbots could probably be safely shared between residents, in the same way that toilets, cutlery, and bedding can be safely shared when properly cleaned.

4.1 Sexbots are morally problematic

The production and use of sexbots is not without controversy, and this controversy pertains to many areas of society – not merely to sexbot use in eldercare institutions. Some of these questions are discussed in more detail in other papers in this anthology; I outline some such arguments here, but there is insufficient space to fully address them all.

It has been suggested that sexbots exacerbate and reinforce gender imbalances, misogyny, sexual exploitation, and the objectification of women (Richardson 2016). The reason for such claims, it seems, it that sex robots (which are generally female in form, and are usually used by heterosexual men) are highly lifelike, and they are of course treated as sex objects. The worry is that legitimising sexbots places us on a slippery slope towards treating or at least viewing real women as sex objects. Even 'generic sexbots' (Lancaster 2021: §2) which are not intended to resemble any person in particular may still encourage users to view women *qua* women as sex objects (and the same may be true of people who use male-appearing sexbots). Moreover, given that most sexbots have "porn star-esque" physiques (Danaher 2017a: 116), this can reinforce uncomfortable notions about the ideal body shape and appearance, in both men and women. Sexbots on sale seem, as far as I can tell, to resemble people aged 16-30[7] (Realdoll 2019, 4woods 2018); the use of such 'young adult' sexbots by elderly people could reinforce derogatory stereotypes such as the 'dirty old man' (Royal College of Nursing 2011: 9) and make elderly people themselves feel uncomfortable. A partial solution could be to create a wider variety of sexbots – ones which appear older, and which have 'imperfections' which make them more similar to real human beings – but we should avoid creating sexbots which resemble particular people who have not consented to their likeness being used (Lancaster 2021). However, creating sexbots which appear more mundane and less like porn stars would still not address feminist concerns (such as those of Richardson 2016, 2019) that sexbots objectify women and encourage misogyny. Although it may be true that the sexbot market *in general* exacerbates these problems, the argument loses its clout when it pertains to sexbot usage in eldercare institutions – this is because most of the people in eldercare institutions are

[7] Some sexbots appear to depict children under the age of 16; these are prohibited under UK law, but even if they were not, such sexbots may be morally problematic (see Danaher 2017b).

heterosexual women, who would therefore choose a sexbot which appears male.

A different problem which could arise from sexbot usage is the concern that 'relationships' with robots are asymmetrical, vacuous, and involve some form of self-deception. Although love robots (or 'erobots') are in development (Anctil/Dubé 2019) any relationship we may have with current sexbots is wholly one-sided (Sparrow 2021, Harvey 2015). Nonetheless, we have the tendency to anthropomorphise things which take humanoid form (Sharkey/Sharkey 2012: 36, Leong/Selinger 2019, Nyholm 2020) – and when these humanlike things also say some endearing phrases, move in sexually enticing ways, and we can engage in sexual intercourse with them, our tendency to anthropomorphise is understandably much greater. In some types of relationship – such as a nurse-patient relationship, it may not matter if a relationship is one-sided (Lancaster 2019, Meacham/Studley 2017) but with sexual and romantic relationships, the need for reciprocity may be felt more keenly. One might suggest that, rather than improving the wellbeing of residents, sexbots provide a false, one-sided, empty experience, whereby lonely old people come to experience unrequited love towards what is essentially an inanimate object. This is a *possibility*, but it may not be a very high-risk possibility; the concern that elderly people will fall in love with a sexbot may simply be unfounded. Even if a lonely elderly person *were* to fall in love with a sexbot, they may nonetheless feel an improved sense of wellbeing even if the love is not reciprocated. The seal-like robot Paro has been shown to improve health and wellbeing in elderly people, reducing stress levels and increasing communication (Tamura et al. 2004, Wada/Shibata 2006) – all this is in spite of the wholly one-sided 'relationship' one has with Paro. If one-sided relationships with robot pets and robot nurses can be beneficial for the people who have them, it is not such a great leap to think that a one-sided relationship with a sexbot may also confer some benefits. At any rate, it would be overly paternalistic to prevent elderly people from having the opportunity to engage with sexbots out of the (perhaps unfounded) fear that they may get too attached to them.

4.2 Elderly people do not want sexbots

It has proved exceptionally difficult to find empirical research into whether elderly people would be willing to use sexbots. Given that many elderly people are so frequently treated as asexual, it is not surprising that there has been so little empirical research into this area. It may well be the case that a substantial number of elderly people – like the younger population – would not want to engage in sexual relations with one of today's sexbots. It may also be the case, however, that as robots become more ubiquitous in other facets of life – for example, carebots, shop assistant robots, and AI lawyers – that people become more accepting of the useful role which robots can play in our lives. This is particularly likely if sufficient advances are made which make sexbots more humanlike in their movements, abilities and 'personalities', and it becomes known that sexbots can provide a satisfying sexual experience.

However, it is worth remembering that even if only a minority of residents would have sex with a sexbot, this does not mean that their desires should be ignored. There does not need to be universal uptake of an activity in order for it to be deemed useful or beneficial to overall quality of life. Activities such as yoga, painting, or singing groups can be recognised as useful and beneficial to quality of life even if only a minority of residents attend the group. This does not, of course, mean that residents have a right to yoga *per se*, but it would be reasonable to claim that residents have a right to physical activity of some description (and yoga is a reasonably good activity, as is aerobics, pilates, tai chi etc). In this paper I am making the argument that residents in care homes have a right to sexual activity / fulfilment *of some description*, and I am suggesting that sexbots are one such option. This is not, however, equivalent to claiming that residents *have a right to sexbots* – just as suggesting that residents have a right to physical activity (and yoga is a useful physical activity) does not commit one to the claim that residents specifically *have a right to yoga*. It is my suggestion then, that some elderly people might want to use sexbots for their sexual fulfilment, and there are few reasons why this should not be permitted.

4.3 We do not expect sexbots

One argument against the provision of sexbots in eldercare institutions is that we do not *expect* them to provide for the sexual needs of their residents, whether through sexbots or any other means. If no one expects sexbots, then why should institutions provide them?

This sort of argument collapses under even the slightest pressure, however, because it is often the case that our expectations are inherently tied to the status quo: we expect eldercare institutions to be the way eldercare institutions *are*. In 1995, we did not expect eldercare institutions to provide free internet access for all residents, but these days, we probably *do* expect it as standard; our expectations have changed as society and care provision has changed. Simply because we do not expect something as standard does not demonstrate that it would not improve residents' lives if it *were* to be provided. We do not expect eldercare institutions to have swimming pools or massage treatments – but residents who live in institutions which have such facilities would probably make use of those facilities, and experience an improved quality of life as a result. If, as I suggest, sexbots have the potential to improve the wellbeing of residents, and residents have a claim to sexual fulfilment as part of their care (as I shall argue shortly), then it would seem to follow that eldercare institutions should consider providing sexbots for their residents. It is true to say that swimming pools, spa treatments and many other facilities could also improve the wellbeing of residents, and it would make sense for eldercare institutions to also consider such facilities – the difference, of course, is that swimming pools and spa treatment rooms are likely to be more pragmatically difficult to provide, whereas sexbots are relatively easy to provide, as are televisions and computers. When something is relatively easy to provide and stands to improve residents' wellbeing, then eldercare institutions should strongly consider their provision.

4.4 Other people do not have constant access to sex

Elderly people (and younger people) who live independently do not all have free and constant access to sex (whether with people or with sexbots), so one might suggest that residents of eldercare institutions cannot reasonably expect to have all their sexual desires accommodated either. Why should residents in eldercare institutions be so privileged as to be provided with something above and beyond what people in everyday society receive?

Although this may seem to be a legitimate argument on first inspection, it is clearly flawed once one reflects on some analogues. Consider: elderly people who live independently may have little or no chance to socialise with others; they may have poor bathroom facilities which cannot accommodate someone who is frail and immobile; they may have unsanitary kitchen facilities and a poor-quality diet; they may be unable to do laundry frequently or effectively, and so may have to wear dirty clothes. The fact that many people living independently can survive in such conditions, does not entail that eldercare institutions need not provide these services. It is reasonable and correct for us to expect that eldercare institutions provide social interaction, clean clothes, a healthy diet, and suchlike. This is primarily because once an institution commits to housing a resident, failing to meet these needs would be failing in their duty of care towards that resident (by contrast, elderly people living alone have no duty of care towards themselves).[8] Additionally, most people have paid for their care (either through their taxes and national insurance contributions, or directly to the institution), and so the provision of adequate care in exchange for that payment is just and right. Institutions should and do provide their residents with clean accommodation and nutritionally balanced meals even though people outside of institutional care might not have such needs met when they are responsible for themselves. Analogously, we can maintain that eldercare institutions have a responsibility to cater for the sexual needs and desires of their residents, even though many people living independently do not have such needs met. This is because sexual gratification is an important part of life, and can improve residents' wellbeing substantially.

It is also worth noting that people outside of institutional care are free to go out and meet other people and *attempt* sexual activity with them in a way that residents of eldercare institutions are not. People living independently have the *capability* (Nussbaum 2011) to seek out relationships, use sex workers, or purchase their own sexbot if they so desire. This means that the sexual needs of people outside of institutional care can be more easily catered for by the person themselves when compared to people within institutional care. We generally consider that people have the right to pursue sexual expression and gratification[9], but care institutions effectively take away this right, by not providing facilities which permit sexual activity, and by limiting residents' freedom. The provision of sexbots within care institutions could therefore help to redress the balance, bringing the capabilities of residents closer to those of people who live independently (see Nussbaum's 2011 capabilities approach – this is developed further in the next section). It would therefore seem reasonable for eldercare institutions to make efforts to cater for people's sexual needs, and providing sexbots is one way of doing this.

[8] Some writers such as Kant (2011, 2017) might disagree, and suggest that we *do* have a duty to care for ourselves adequately; I am unconvinced about this, but even if it were true, this is not the same as the legal duty of care which eldercare institutions have towards their residents.

[9] Assuming, of course, that this does not infringe on others' rights.

4.5 Sex is a want, not a need

I have at various points in this paper made reference to 'sexual needs', but an opponent may wish to argue that sex is a want rather than a need. Although there have been suggestions (see Maslow 1943) that sex *is* a physiological need, along with air, food, water, and sleep, it is nonetheless true that one can live well into their hundreds having never engaged in sexual activity, whereas one cannot live for very long at all without air, food, water, and sleep. Even people who used to be very sexually active can survive for many years in institutional care without sex, so in that sense sex does not seem to be a *need*.

Even though it seems true that lack of sex is not a threat to life in the same way that lack of food is, it nonetheless seems evident that sex is an important part of wellbeing. Martha Nussbaum suggests that welfare is inherently linked to the capability of people to be or do particular things. One of the central capabilities she identifies is bodily integrity, which includes "having opportunities for sexual satisfaction" (Nussbaum 2011: 33). Her argument is that societies should support this capability (among others) to safeguard the welfare of their citizens. The key feature of Nussbaum's capabilities approach is that people should be given *opportunities* to fulfil the capabilities she lists – whether or not people actually make use of the opportunity is their choice, but having the opportunity is in itself valuable.

For many people, a life *with* sex is very much preferable to a life *without* sex. The same argument can be made for conversation, friendship, leisure activities, and entertainment. Living without friendship and suchlike will not cause a person's death, but it will almost certainly make life very miserable; life *with* friendship is a lot better than life without it. I suggest that the same is true of sexual activity. One does not have to demonstrate that sex is a physiological necessity; it is sufficient to point to the simple fact that that sex improves one's quality of life (Hajjar/Kamel 2003, Gott/Hinchcliff 2003, Royal College of Nursing 2011: 7), as do friendship and leisure activities. As it happens, sex also improves one's physical health and life expectancy (Hajjar/Kamel 2003), but even if it were only the case that sex improved subjective perception of quality of life, that could still be sufficient reason to encourage institutions to facilitate its occurence.

4.6 People cannot have everything they want

One might argue that simply because something would improve quality of life does not mean that eldercare institutions simply *must* provide it. As I noted above, a swimming pool would be a great facility if it were to exist in an eldercare institution, and would improve residents' lives, but we cannot therefore demand that all such institutions get a swimming pool. There are a great many things which elderly people may desire, and which could improve their quality of life, but which are simply not conducive to life within institutional care. Examples include the chance to be around animals, attend concerts, visit the beach every day, live with young children, keep exotic pets, attend rallies, travel abroad, and any number of other activities. I mentioned above that one of Nussbaum's central capabilities is bodily integrity – part of this includes "being able to move freely from place to place" (Nussbaum 2011: 33) but clearly, care homes limit this capability. Residents are not free to go to the beach, travel abroad, and attend rallies whenever they choose, yet we do not suggest that care home staff should cater for this desire by taking residents wherever they like, whenever they like. The financial cost and logistical difficulties involved make them untenable. Although the above activities are fun and can improve wellbeing, they are not pragmatically feasible for residents in eldercare institutions; someone may suggest that the same is true of sexual activity.

However, I believe that such a suggestion would be misguided. Allowing for sexual expression would not require a great deal of change for an institution (unlike the other examples given above, which would require extensive risk assessments and would be logistically difficult). In order to cater for sexual activity with a sexbot, institutions would merely need to provide some privacy (such as lockable doors or a dedicated room), and a sexbot which is thoroughly cleaned and up to the task (in section 5, I address how sexbots would need to change in order to be useful for elderly people). Even initiatives such as nurses knocking and waiting before entering a resident's room could be enough to provide ample privacy for sexbot usage. Eldercare institutions would thus be able to easily cater for residents who wish to use a sexbot. Although it is fair to argue that people in eldercare institutions cannot have *everything* they want, when a facility can be easily implemented and would provide a sufficient improvement to people's quality of life, the reasons not to provide it seem trivial. The fact that something would be an inconvenience for care providers is not a sufficient reason not to provide it (Sharkey/Sharkey 2012: 37).

Questions of whether residents in eldercare institutions should have access to sexbots is in many ways a question of social justice. John Rawls famously suggested that justice is "the first virtue of social institutions" (Rawls 1999: 3). Given that money and other resources in care homes are scarce commodities, the costs (both financial and socio-ethical) and benefits of providing sexbots need to be carefully weighed. It may not *always* be viable for care institutions to provide sexbots for elderly residents to use, but I suggest that their provision should at the very least be a consideration, given the potential boost to residents' wellbeing.

Thus, I suggest that sexbots can and should be provided in eldercare institutions. To be clear, my argument is not that sexbots are the *only* solution to the problem of satiating the sexual desires of elderly people, nor even that sexbots are necessarily the *best* solution for everybody. I am simply arguing that sexbots are a plausible solution that could help residents who still have sexual drives but are currently unable to find a suitable outlet for them.

5 Are sexbots up to the task?

I believe I have made a convincing case in favour of providing sexbots for elderly people in care institutions. I shall now proceed with addressing how sexbots could become capable of fulfilling the purpose I am suggesting that they could fulfil – in other words, what needs to happen in order for sexbots to be up to the task. Unfortunately, it seems that today's sexbots could not be used effectively for sexual intercourse with elderly people because they are heavy, fairly inert, and not dextrous enough. These issues could cause two types of problem:

a) Elderly people getting injured.
b) Elderly people being unable to engage in (satisfying) sex with a sexbot.

The first concern echoes a prominent concern regarding the use of robots of all types (including carebots, robot nannies, military robots, and driverless cars): we worry that robots may malfunction and hurt people. Given that elderly people in institutional care may be frail or infirm, there is a very real possibility that if robots malfunctioned, the elderly people could suffer serious injuries.

However, this alone is not a convincing argument against the use of sexbots, because there are many technological devices which are used in eldercare institutions which also have the potential to injure residents if they malfunction. For example, motorised tilting chairs (to assist in standing and sitting), stairlifts, equipment to lift patients in and out of bed, motorised wheelchairs, adjustable beds, and exercise machines – not to mention medical equipment such as ventilators – all have the potential to seriously harm an elderly person if they malfunction.

The response, however, is not to avoid such equipment altogether, but rather, to ensure that the devices are equipped with safety features which prevent malfunctions and/or that harm is minimised in the event of a malfunction. There is no reason why sexbots for elderly people cannot be fitted with failsafe functions too. For example, they could be able to alert a nurse if: a 'panic button' is pressed; a user says a trigger word (such as "Help!"); a user has not moved for a period of time; or the sexbot detects a malfunction in itself. These sorts of functions would help to maintain privacy, whilst increasing safety. Sexbots could also be fixed in particular ways so as to prevent them from falling. Someone might suggest that sexbots are inessential and so are not worth the safety risk, whereas the other equipment used in care homes is essential. However, this is not entirely true. A nurse could push a resident in a wheelchair or manually pull them out of a chair or bed, rather than using motorised equipment to accomplish these tasks. Motorised wheelchairs, hoists, and tilting chairs are not essential equipment, yet the benefits gained from them are substantial enough to warrant using them, in spite of the risk of malfunction. The same may be said of sexbots.

What is perhaps a more pressing and legitimate concern is that elderly people may get injured during sex with a sexbot because of the weight and rigidity of a sexbot, even if it does not malfunction. Generally, when someone has sex with a sexbot, the sexbot lies still while the human is on top and doing all the work – something which may not be safe or possible for a majority of frail elderly people in care institutions.[10] Lying underneath the sexbot is not a viable option either, because sexbots are so heavy: female sexbots are generally around 30-60kg (Smart Doll World 2021) – and male sexbots weigh

[10] Of course, if some elderly people *can* take an active role in sex with a sexbot, then they would not need these modifications.

a little more – and cannot support their own weight on their hands or legs as a human would during sexual intercourse.

This means that even if a sexbot is functioning normally, there is still a genuine risk that a frail or infirm elderly person could be crushed or injured by the sheer weight of it. Even if a resident *could* withstand the weight of a sexbot lying on top of them, this would not be sufficient to facilitate intercourse, because the sexbots currently available are so passive during sex. If a person were to attempt sex with the sexbot on top of them, they would have to lift or manoeuvre the sexbot repeatedly in the appropriate ways so as to achieve the feeling of sex – something which would require great upper body strength, dexterity, and stamina. This precludes frail, elderly people from being underneath during intercourse with a sexbot.

This means that today's sexbots are unable to perform the very task that I am arguing they should be utilised for, which may seem absurd to some readers. However, advances in robotic technologies can happen quickly, and it would not take too great a technological development to create a sexbot which is better suited to elderly users. A sexbot for elderly people would need to be:

a) Lighter weight
b) Able to support its own weight on its hands (or in some other way)
c) Gentle with its user
d) Able to take a more active role in sex

Technological convergence is a process whereby previously separate forms of technology merge into a single technology. This has happened with mobile phones, which can now function as a sat nav, games console, alarm clock, and camcorder, all of which used to be separate technologies. Roboticists also incorporate previously separate technologies into new models of sexbots, in order to create new sexbots which are more advanced than their predecessors. For example, whereas older sexbots were little more than dolls, some of today's sexbots have AI components which better enable them to engage in 'loving' conversations, making them not just sexbots, but lovebots too (Anctil/Dubé 2019).

A sexbot with the four modifications I suggest above could easily fulfil the brief I am outlining here, and so even if we are currently at a technological juncture where sexbot technology is unable to meet the brief, my argument can still stand. When adequate sexbots exist, I suggest that they can and should be provided for residents in eldercare institutions.

6 Conclusion

During recent decades, significant progress has been made in providing for the social and emotional welfare of elderly people in institutional care; however, sexual activity and sexual pleasure for elderly people still remains something of a taboo. Nonetheless, research has shown that elderly people still have sexual desires, and that a life with sexual pleasure is preferable to a life without it. I have argued herein that elderly people in institutional care should have their sexual needs catered for, and that sexbots would be a useful solution to the problem. I am not suggesting that sexbots should be used exclusively and that other sexual outlets (such as masturbation, sex workers, or relationships with other residents) should be closed off; rather, I suggest that sexbots should be provided as one possible sexual outlet among many.

Sexbots in their current form are generally designed for able-bodied heterosexual men, and are therefore not well-suited to providing sexual pleasure for frail, elderly people – most of whom are women – with limited dexterity. Technologies do exist, however, which could potentially be incorporated into sexbot design to enable the most vulnerable of people to have a satisfying sexual experience. It is my suggestion that sexbot developers should make these changes to their sexbots, and that such sexbots should be provided in eldercare institutions, so that residents can continue to explore their sexuality for as many years as they wish to.

References

4woods (2018): "Real Love Doll" 4woods, September 17, 2019; https://aidoll.4woods.jp/en/.

Age UK (2020): "Safeguarding Older People from Abuse and Neglect"; https://www.ageuk.org.uk/globalassets/age-uk/documents/factsheets/fs78_safeguarding_older_people_from_abuse_fcs.pdf.

Age UK (2021): "Protection from Abuse"; https://www.ageuk.org.uk/information-advice/health-wellbeing/relationships-family/protection-from-abuse/.

Anctil, D./Dubé, S. (2019): "Beyond Sex Robots: Erobotics Explores Erotic Human-Machine Interactions.", Phys Org., https://innerself.com/personal/relationships/couples/sexuality/20892-beyond-sex-robots-erobotics-explores-erotic-human-machine-interactions.html.

British Geriatrics Society (2018): "Depression among Older People Living in Care Homes Report", Royal College of Psychiatrists,; https://www.bgs.org.uk/sites/default/files/content/resources/files/2018-09-19/Depression%20among%20older%20people%20living%20in%20care%20homes%20report%202018_0.pdf.

Care Quality Commission (2020): "Promoting Sexual Safety through Empowerment"; https://www.cqc.org.uk/sites/default/files/20200225_sexual_safety_sexuality.pdf.

Chevarie-Cossette, S.-P. (2017): "Prostitution: You Can't Have Your Cake and Sell It." In: Journal of Practical Ethics 5/2, pp. 77-84.

Crown Prosecution Service (2019): "Prostitution and Exploitation of Prostitution"; https://www.cps.gov.uk/legal-guidance/prostitution-and-exploitation-prostitution.

Danaher (2017a) "The Symbolic-Consequences Argument in the Sex Robot Debate". In: J. Danaher/N. McArthur, Robot Sex: Social and Ethical Implications, Cambridge, MA: MIT Press, pp. 103-132.

Danaher (2017b) "Robotic Rape and Robotic Child Sexual Abuse: Should They be Criminalised?". In: Criminal Law and Philosophy 11/1, pp. 71-95

Di Nucci, E. (2017): "Sex Robots and the Rights of the Disabled". In: J. Danaher/N. McArthur, Robot Sex: Social and Ethical Implications, Cambridge, MA: MIT Press, pp. 73-88

Franowski, A.C./Clark, L.J. (2009): "Sexuality and Intimacy in Assisted Living: Residents' Perspectives and Experiences." In: Sexuality Research and Social Policy 6, pp. 25-37.

Gott, M./Hinchcliff, S. (2003): "How Important Is Sex in Later Life? The Views of Older People". In: Social Science and Medicine 56, pp. 1617-1628.

Hajjar, R./Kamel, H. (2003): "Sexuality in the Nursing Home, Part 1: Attitudes and Barriers to Sexual Expression". In: Journal of the American Medical Directors Association 4, pp. 152-156

Harvey, C. (2015) "Sex Robots and Solipsism". Philosophy in the Contemporary World 22/2, 80–93.

Ismail, S. (2013): "Brothels for Disabled People: Guess What? We like Sex Too!" In: The Independent [online] 23 January; https://www.independent.co.u k/voices/comment/brothels-for-disabled-people-guess-what-we-like-sex-too-8461537.html

Kalra, G./Subramanyam, A./Pinto, C. (2011): "Sexuality: Desire, Activity and Intimacy in the Elderly". In: Indian Journal of Psychiatry 53/4, pp. 300-306.

Kant, I. (2011): [1785]: Groundwork of the Metaphysics of Morals,ed. by Gregor, M./J. Timmerman, Cambridge University Press; https://r2.vlereader.com/Reader?ean=9781139006798.

Kant, I. (2017): [1797]: Metaphysics of Morals. ed. by Denis, L., Cambridge: Cambridge University Press.

Lancaster, K. (2021): "Non-Consensual Personified Sexbots: An Intrinsic Wrong". In: Ethics and Information Technology 23/4 pp. 589-600.

Lancaster, K. (2019): "The Robotic Touch: Why There is No Good Reason to Prefer Human Nurses to Carebots". In: Philosophy in the Contemporary World 25/2, pp. 88-109.

Leong, B./Selinger, E. (2019): "Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism". In: Proceedings of the Conference on Fairness, Accountability, and Transparency, held 2019 at Atlanta, GA, USA. ACM Press, 299–308, http://dl.acm.org/citation.cfm?doid=3287560.3287591.

Little, R. (2007): "Hospice Finds Prostitute for Disabled Man". In: Oxford Mail [online] January 26. ; https://www.oxfordmail.co.uk/news/1149422.hospice-finds-prostitute-disabled-man/.

Maslow, A. (1943): "A Theory of Human Motivation." In: Psychological Review 50/4, pp. 370-396.

Meacham, D./Studley, M. (2017): "Could a Robot Care? It's All in the Movement". In: Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence, Oxford University Press, 97–112.

National Center on Elder Abuse (2018): "Sexual Abuse in Nursing Homes: What You Need to Know." In: National Consumer Voice for Quality Long-Term Care. October 6, 2021; https://eldermistreatment.usc.edu/wp-content/uploads/2018/09/Consumer-Voice_sexual-abuse-issue-brief-FINAL.pdf.

Nussbaum, M.C. (2011): Creating Capabilities: The Human Development Approach, Cambridge, MA: Harvard University Press.

Nyholm, S. (2020): "Humans and Robots: Ethics, Agency and Anthropomorphism". Rowman & Littlefield.

Office for National Statistics (2014): "Changes in the Older Resident Care Home Population between 2001 and 2011 - Office for National Statistics", September 28, 2021; https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/articles/changesintheolderresidentcarehomepopulationbetween2001and2011/2014-08-01.

Public Health England (2015) Health Promotion for Sexual and Reproductive Health and HIV: Strategic Action Plan 2016 to 2019,; https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/488090/SRHandHIVStrategicPlan_211215.pdf.

Rawls, J. (1999): A Theory of Justice: Harvard University Press.

Realdoll (2019): "Build Your Realdoll", September 14, 2019; https://www.realdoll.com/product/build-your-realdoll/.

Realdoll (2021): "Real Dolls in Stock"; https://www.realdoll4me.com/en/products/real-dolls/.

Richardson, K. (2016): "The Asymmetrical "Relationship": Parallels Between Prostitution and the Development of Sex Robots." In: SIGCAS Computers & Society 45/3, pp. 290-293

Richardson, K. (2019): "Sex Robots: The End of Love". Polity Press.

Roach, S.M. (2004): "Sexual Behaviour of Nursing Home Residents: Staff Perceptions and Responses." In: Journal of Advanced Nursing 48, pp. 371-379.

Royal College of Nursing (2011): "Older People in Care Homes: Sex, Sexuality and Intimate Relationships", November 30, 2018; https://www.equalityhumanrights.com/sites/default/files/004136.pdf.

Settegast, S. (2018): "Prostitution and the Good of Sex." Social Theory and Practice 44/3, pp. 377-403.

Sharkey, A./Sharkey, N. (2012): "Granny and the Robots: Ethical Issues in Robot Care for the Elderly." In: Ethics and Information Technology 14/1, pp. 27-40.

Shenzhen All Intelligent Technology Co. Ltd (2022): "Artificial Intelligent Sex Robot"; https://sexrobot.en.alibaba.com/.

Smart Doll World (2021): "Weight", October 5, 2021; https://www.smartdollworld.com/product-category/weight/.

Smart Doll World (2022): "AI Robot Sex Dolls"; https://www.smartdollworld.com/ai-robot-sex-dolls/.

Social Care Institute for Excellence (2006): "Assessing the Mental Health Needs of Older People - Depression", September 27, 2021; https://www.scie.org.uk/publications/guides/guide03/problems/depression.asp

Sparrow, R. (2021): "Virtue and Vice in Our Relationships with Robots: Is There an Asymmetry and How Might It Be Explained?' International Journal of Social Robotics 13, 23–29.

Spector, J. (ed.) (2006): Prostitution and Pornography: Philosophical Debate About the Sex Industry, Stanford University Press

Stepler, R. (2016): "World's Centenarian Population Projected to Grow Eightfold by 2050",; https://www.pewresearch.org/fact-tank/2016/04/21/worlds-centenarian-population-projected-to-grow-eightfold-by-2050/.

Stretch, E. (2013): "Special Visits: Care Home Admits Inviting Prostitutes to Have Sex with Disabled Residents". The Mirror [online] January 28.; http://www.mirror.co.uk/news/uk-news/prostitutes-invited-into-care-home-for-sex-1560860.

Tamura, T./Yonemitsu, S./Itoh, A./Oikawa, D./Kawakami, A./Higashi, Y. (2004): "Is an Entertainment Robot Useful in the Care of Elderly People with Severe Dementia?". In: Journals of Gerontology Series A Biological Sciences and Medical Sciences 59, 83–85.

Tarzia, L./Fetherstonhaugh, D./Bauer, M. (2012): "Dementia, Sexuality and Consent in Residential Aged Care Facilities". In: Journal of Medical Ethics 38/10, pp. 609-613.

Thomsen, F.K. (2015): "Prostitution, Disability and Prohibition". In: Journal of Medical Ethics 41/6, pp. 451-459.

TLC Trust (2021): "Responsible Sexual Services for Disabled People", September 29, 2021; https://tlc-trust.org.uk/.

Vandervort, L. (1987): "Mistake of Law and Sexual Assault: Consent and Mens Rea". In: Canadian Journal of Women and the Law 2/2, pp. 233-309.

Varden, H. (2006): "A Kantian Conception of Rightful Sexual Relations: Sex, (Gay) Marriage and Prostitution." In: Social Philosophy Today 22, pp. 199-218

Vicente, A. (2016): "Prostitution and the Ideal State: A Defense of a Policy of Vigilance". In: Ethical Theory and Moral Practice 19/2, pp. 475-487.

Wada, K./Shibata, T. (2008): "Social and Physiological Influences of Robot Therapy in a Care House". In: Interaction Studies 9/2, 258–276.

Westin, A. (2014): "The Harms of Prostitution: Critiquing Moen's Argument of No-Harm". In: Journal of Medical Ethics 40/2, pp. 86-87.

WHO (2018): "Ageing and Health", March 28, 2019; https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

Alice Does not Care
Or: Why it Matters That Robots "Don't Give a Damn"

Imke von Maur

1 Introduction

In his documentary *Alice cares* (2015), Sander Burger presents a pilot project of researchers in Amsterdam in which a robot called Alice is introduced as a new friend to three elderly women. In this text, I will raise serious doubts that Alice, as the title of the movie claims, cares. On the contrary, I point out why Alice does not and in principle cannot care. This circumstance is illustrative of John Haugeland's utterance with regard to the general problem of artificial intelligence (AI) – namely that computers "don't give a damn". All that robots do is pattern recognition and giving a seemingly adequate output to a given input. If the capacity to discriminate between *this* and actual care gets lost, as I argue, there is the danger that not only the concept of *care* changes, but ultimately the practice of caring.

Alice is a paradigmatic example of this and highlights, on the one hand, the structural presuppositions that lead to the development of such technology in the first place, and, on the other hand, the severe consequences this has for society and our understanding of what it means to be human and to live a good life. Taking seriously the quintessence of this text comes with consequences for the research and implementation of so-called social robots, which are supposed to care not only for elderly people but be companions of lonely teenagers and adults in times of a pandemic.

I will argue that implementing robots in order to care and to reduce loneliness not only fails – for robots simply cannot care – but has the potential to make things worse in a systematic way. An ethical assessment of care robots thus should not only tackle the concrete outcomes of a care robot in a given scenario – I call this a *functionalist-individualist approach* – but needs to be concerned with the broader socio-political structural presuppositions and effects.

One *structural effect* of normalizing the implementation of care robots could be that humans will not only stay lonely while interacting with robots, but that they will give up expectations of real care and true relationships in the first place. This might lead to a severe change in (expectations within and about) social interactions at a large scale. The effects a continuous interaction with a care robot might have on the affective and social repertoire of a person thus exceeds the scope of this individual. An example for a *presupposition* of normalizing simulated care is the techno-solutionist narrative, according to which societal problems can be satisfactorily dealt with by implementing technological solutions. This narrative makes intelligible the implementation of care robots to researchers, tech-companies, the public, political decision-makers and ultimately the customer in the first place. There are other potential imaginations and narratives on how to provide care which should be considered in an ethical assessment of care robots.

The aim of this paper is twofold: Firstly, I demonstrate why care robots do not and in principle cannot care. This is a rather analytical point. Connecting to this, I make the normative argument that it in fact matters that humans *actually* care about another instead of *simulating* to do so. This normative step provides the ground for my second aim in this paper, namely to draw attention to the necessity of a larger socio-political approach for an ethics of care robots (see also Coeckelbergh 2022 who recently argues for a "political philosophy of AI"). That is, I aim to provide the grounds for a different focus for assessing the appropriateness of using care robots to reduce loneliness which goes beyond local-individualistic and functionalist perspectives but considers the broader socio-political horizon. It is this structural perspective I am concerned with throughout the text. Thus, it is important to clarify that I am not arguing at the level of concrete individuals interacting with robots. The "ethical subject" I address in this text is not a concrete subject but society at large in its responsibility for how discourses are shaped and for which narratives become powerful. The caveat being made here thus concerns the temptation to take what I argue for as prescriptive suggestions for prohibitions on the individual level. That is not the goal. The goal is to raise awareness to the potential structural consequences of the wrong assumption that robots could care about us. That is, consequences for societally shared narratives and imaginations about how to live a good life (together) and on related practices and institutionalizations, among others.

2 Alice – "The care of tomorrow"?

> *Alice:* What makes you happy?
> *Ms. Remkes:* I haven't figured out yet. [pause] "What makes you happy?" [laughs seemingly ashamed and looks away] I have to think about that.
> *Alice:* What else could I do for you?
> *Ms. Remkes:* [face starts to get angry] I don't feel like having a robot in my home. [looks around with a dissatisfied face] I prefer a living human being.
> *Alice:* Oh, that's a shame. Thank you for the conversation. Maybe I'll see you soon.
> *Ms. Remkes:* We'll wait and see.

The first version of Alice that is used in the documentary is a small doll with a puppet-like female face that should feel soft and skin-like. The body, though, is that of a plastic toy and does not resemble a human, although there is a torso, two arms and two legs. Alice is small, it can only sit on the couch or a chair like a doll but not move itself or walk around, let alone do any physical task for those it is supposed to care about. The 'caring' is meant to be one of companionship: "Alice's first mission is to assume a social role and be treated as a social entity. That is why we designed a social robot that is both, a friend as well as a connector to others. (Alice Cares Promo 2021)" On the webpage "Alice cares", its main selling point is announced as its ability to "decrease feelings of loneliness": By engaging in "a conversation, see[ing] what her companion is doing and by scanning her environment, Alice is able to react and adapt to situations in her surroundings." (ibid.) The robot is furthermore said to remember things its counterpart said, to ask them questions, and to motivate them to engage in specific activities like going for a walk, singing a song or writing a letter.

The documentary encourages the audience to be skeptical at the beginning of the story since the three women are initially not portrayed to be enthusiastic about the robot. But, over the course of the story, the subjects seem to become more and more attuned to the robot, to accept it and even to like it, to engage in conversations with it and ultimately find a companion in it. I interpret this narrative of a *seemingly* stepwise acceptance to be manipulating the viewers of the documentary and think it obscures a serious engagement with both the problem and the three women.[1] One scene at the end of the documentary shows how Ms. Remkes goes to a café with Alice (the scientist carries Alice to this place for it cannot walk). Ms. Remkes seems to be proud of being there with Alice and talks to the waitress about her being part of the research and pilot project. This scene is meant to elicit the feeling in the viewer that Ms. Remkes is happy because of Alice being her companion. But this is a typical instance of the "Hawthorne-effect": The circumstance that she is part of a study changes her feelings and behavior.[2] Ms. Remkes gets attention, people are aware of her because of the camera team, of the robot and the special situation. She is *recognized* by other *humans* as being special, important, ultimately: *as being there*. Another misleading thought the documentary suggests is that, even though Alice might not completely compensate for human interaction, it is at least sophisticated entertainment, like an advanced TV, one that supposedly reacts and responds instead of being merely looked at. There seems to be an intuitive suggestion to say that this is better than nothing. But nothing is not the only alternative. The problem is already framed in a way in which it seems to be the only available and thus inevitable solution to rely on technological progress in order to counter societal problems and developments. The *crucial* question, that motivates the following considerations is: How is it that researchers, political decision-makers and the broader public really consider it possible for a machine to care?

In the following section, I shed light on the complexity and depth of the phenomenon of care that I take to be indispensable for humans and that (not

[1] I will sketch some indicators for this interpretation in what follows without the aim to generalize this critique by inferring that *all* documentaries about care robots *necessarily* have to work this way. Also, the manipulative character of the documentary alone does not prove that the elderly women portrayed or even any elderly person engaging with a care robot has to be unhappy with that. What I aim at with this opening here is to highlight the framing of one contribution for a discourse about a normative assessment of care robots – namely this specific documentary – which I find problematically suggestive.

[2] From 1927 to 1932 researchers conducted a field study in the Hawthorne department of the American Western Electric company to figure out, which factors would increase the productivity of the workers (Roethlisberger & Dickson 1939). What is now known as the "Hawthorne effect" is the influence of the very setting of a study on the attitude and behavior of the subjects under investigation (Sanders & Kianty 2006: 59ff.). This phenomenon, among other similar effects, is also investigated under the term "reactivity", for instance in a recent and encompassing research project by Marion Godman, Caterina Marchionni and Julie Zahle (Reactivity Project 2021) .

only) individuals in the case study of Alice miss. Afterwards, I reveal the understanding of care underlying care robots and show that the two have nothing to do with each other, i.e., that Alice does not care.[3]

3 Caring: Being (taken as) a person

Care is a complex and manifold phenomenon. To care means, first and foremost, to not be indifferent to what happens to another. This is not bound to the necessity that the other is of high personal import to me like friends or beloved ones, not even that a teacher or nurse for instance has to like all their students or patients wholeheartedly. But it manifests differently in caring bodies what happens to the one they *care about*. For the topic of the paper, I am concerned with *caring about* as "a mental capacity or a subjective state of concern" (van Wynsberghe, 2013: 414) and not *caring for* meant as "an activity for safeguarding the interests of the patient" (ibid.). While I believe the latter is not properly delivered without the attitude being at issue in the former one, an argument for this is not the topic of debate here. As I want to provide the ground for an argument against emotional bonds with robots designed to be social *companions* – it cannot be meant that "care" is only the performance of tasks (*care for*) like carrying patients to bed or reminding them to take their medicine, but must include an *attitude*, by which a companion is characterized (*care about*).[4] It is not simply "caregiving" in the sense of sustenance but care in the sense of a meaningful relationship I am concerned with in the following.[5]

[3] In this paper, I am concerned with robots designed to be companions and to reduce loneliness. There are artificial systems called robots or care assistants as the famous seal Paro used for patients with severe dementia, which might work as sophisticated entertainment but are not intended to be a companion who cares. What I discuss in this paper is not about Paro or robot dogs or other such devices.

[4] "In the field of care ethics, Joan Tronto claims that good care is the result of both a caring attitude in combination with a caring activity (Tronto 1993). In other words, a marriage between the dimensions of caring about and caring for." (van Wynsberghe 2013: 414)

[5] I will solely be concerned with human-human relationships in this section and will remain silent about the possibilities of care-relationships between humans and animals or plants.

In this sense, as said, a caring person is *affectively not indifferent* to what is going on. This is self-explanatory for intimate relations characterized by a strong emotional bond – loving relations, friendships, family relationships etc. But also, in a relationship of care in the medical or educational context the involved subjects react affectively, depending on the other's situation. This becomes especially clear when concrete encounters are considered. In situations where humans engage with each other face-to-face, they look at each other, recognize vocalizations and facial expressions and their meaning in this specific context against the background of a *shared space of meaning*. Humans and their encounters are inevitably situated in concrete practice-specific contexts, in which something is "at issue and at stake" (Rouse 2002) that is often hard to put into words. A person is not just sad. A concrete human being with concrete needs, desires and hopes, with a concrete and specific (affective) biography makes sense of their situation (affectively).[6]

To care about this person means to bring them into existence not only *as a person* (rather than an object) but as *this very concrete individual*.[7] A caring person brings into existence the other one as a valuable being, a concrete individual with specific needs, concerns, beliefs, values and so on. Phenomeno-

[6] Elsewhere I elaborate on the human capacity to (tacitly) navigate in such spaces by adopting a practice theoretical and phenomenological approach to the disclosure of meaning, which is not only characterized by affectivity but by sociality and the knowledge about relevant rules of practice specific "games", as Bourdieu would calls this (von Maur 2018 and 2021). The tacit (social) knowledge humans incorporate goes beyond what is implementable in a machine, inter alia because there are too many and too subtle things to be known in order to skillfully engage in such domain specifically meaningful "little worlds" (ibid.).

[7] The claim is not meant in the sense that the human being would not exist without this act of recognition. For sure, the flesh of a creature does not fade away if nobody recognizes it as a person. But as Judith Butler claims with regards to the question the life of whom is grievable, the ontological, epistemological and ethical aspects of "apprehending a life" are inextricably intertwined (2009: 2ff.). If persons are not apprehended as persons, not *brought into existence* as vulnerable beings, they do not appear in our repertoire of the ones we grieve for. Butler is especially interested in the framings, for instance of media, which determine the life of whom is apprehended and of whom not. I take it to be worth further research to apply her framework – which in her book is applied to war and its victims – to the field of the elderly and children. These societal groups are especially vulnerable and deserve a cautious consideration and treatment. Yet, often they do not seem to be apprehended in a way necessary to initiate appropriate conditions for care and education. Children and the elderly rather seem to be effortful and obstacles for the working subject in a neoliberal society.

logically this is nicely illustrated by philosophers like Jean-Paul Sartre or Emmanuel Levinas who emphasize the import of concrete encounters for ethics, i.e. by pointing out what it means to look our counterpart into their "Antlitz" (Levinas). This is an act of personification, an act where we recognize the other one as another ethical subject, a subject with unimpeachable human dignity and human rights. Martin Buber (1973) speaks of "making the other one present" [*personale Vergegenwärtigung*] – meaning to perceive the other one as a whole and as being unique – and already 50 years ago criticized that in modern times there is an omnipresent analytical and reductive perception of others rather than such a perception of the other one as a concrete valuable being (1973: 285). If a person is not able to recognize the other person in the way just described, if they are not able to be affected, it is hard to see how they can care. To care about someone thus means to engage in a relationship and to make the other person matter, to decide that their concerns and well-being matter to oneself – ultimately: to make the concerns of another person one's own concerns.

This does not imply that to care means to sacrifice. There is no universal or personal duty involved in caring. Instead, to care presupposes that the caring person *decides autonomically* to care about the other, to take part in their life.[8] Although this is quite demanding, people feel existentially threatened if they imagine that *nobody* cares about them *in this way*. For sure, the degree in a loving relationship or intimate friendship is much higher when it comes to this feature of relationships, but, still, this is what makes up other caring relationships as well – i.e., educational or medical ones. Crucially, in any caring relationship one *responds* to the other as well as taking *responsibility* for them.

Although taking the counterpart seriously without it being a sacrifice or involving patronizing attitudes is key for caring relations, there are certainly differences in the kind of (a)symmetry of caring relations. To care about children who have not developed a decisive degree of autonomy yet, or to care about an elderly person who has lost this autonomy (to a certain degree), are

[8] This can be nicely illustrated by help of Tzvetan Todorov's (1993 [1991]: 80-101) distinction between care and mercy (and solidarity or charity). The merciful person sacrifices and assumes a burden, whereas the caring person just cares. This does not mean that this is not effortful or experienced painfully. I cannot but care about my children although this is often demanding and painful, and in the same manner, real care in the medical or educational or other contexts, implies the attitude of making the problems of the other one's own – not in the intensity of a parent-infant relation, but in a way that goes beyond mercy.

instances of rather asymmetrical relations. The one who cares is the one who takes the responsibility for the less autonomous member of the relationship. Yet, it is key to *take seriously* the care-receiving person *as a person* and to take seriously that there is, also for a less autonomous person, like a child or elderly person, a possibility to "care back".[9] Thus, also the one who cares is a concrete counterpart, an individual speaking with their own voice, one who can also disagree with the other, be angry at them or be in need of consolation. If a neighbor, relative or nurse cares for and about an elderly person, the care-giver plays a role in their life. At least in such a relationship of care that lasts for a longer period, the elderly individual also includes the care-giver in their thoughts and considerations, takes an interest in their well-being, achievements or sufferings. Phenomenologically, in a concrete situation there are two parts who can establish *resonance* (in the sense of Hartmut Rosa 2016). That is, there is not a mere echo of one's own voice but a being who can understand, misunderstand, like or dislike me, who can be similar in some regards rather than others and the engagement with whom can feel comfortable – I can reach the other one in Rosa's sense – or it feels alienating because I feel misunderstood. Both parts in such a relationship that is phenomenologically experienced in specific ways need to be taken as fallible, but first and foremost unconditionally valuable individuals – independently of whether the relationship is (always) experienced as harmonic or rather discordantly.

It is important to keep in mind that I am not claiming that any human professional care-giver could or should care about any of their patients at any time in an emotionally exhaustive way. A certain degree of "professional" (affective) distancing seems to be crucial for professional care-givers in order to stay healthy and not to exploit their (affective) resources. The characterization of care just sketched is meant to highlight crucial aspects of the *attitude* of caring about another *in general* – as a state of *concern* that is characterized by not being indifferent. To care (in the sense of the German word "Sorge" rather than "Pflege") is an affective attitude and practice that provides the basis or

[9] I thank the Research seminar of the Centre for Ethics in Pardubice for discussing this point with me. The idea that to be *cared about* implies the possibility to *care back* makes clear that not to be lonely also means that one is able to partake in the life of another one – namely to worry about their well-being as well. Yet, this is not possible for *any* person who still might be truly cared about – think for instance about patients with severe dementia or people in vigil coma. I do not claim that one cannot truly care about these people who cannot care back in the sense sketched.

background of a relationship. A professional care-giver can care in this sense without necessarily being highly emotionally involved in any interaction with their patients. The crucial aspect rather is the real interest they take in the well-being of the other one, that it makes a difference to them whether the patient is happy or sad, is in a good or bad bodily condition, suffers from social isolation or shares parts of their life with others.

Before moving on to explain why a robot does and cannot care in the way described here, I illustrate the concept of care that in fact *is* at issue in so-called care or social robots – i.e., an atomistic approach to care that reduces the complex meaningful phenomenon just described, to input-output-processes in a functionalistic manner. This 'care' implemented in robots has nothing to do with *caring about,* as I described it. Yet, normalizing the use of the concept 'care' for denoting simulated care involves the danger of powerfully disempowering narratives and imaginations about what it means to entertain a relationship, what it means to care and to be cared about and ultimately, what it means to live a good life.

4 Reductionistic encounters: A functionalist understanding of care

"The trouble with artificial intelligence is that computers don't give a damn". This utterance of John Haugeland (1998: 47) highlights the thesis that robots cannot care in the sense sketched above. Underlying the incapacity to "give a damn" is the robots' incapacity to be affected. I take for granted in this paper that robots do not have emotions, that they lack consciousness and thus any kind of first personal experience in terms of "what it is like" (Jackson 1982).[10] The difference between so-called strong and weak AI (Searle 1980) does not only apply for cognitive but also for affective processes. The thesis of weak AI, namely that artificial systems are able to *recognize* and *simulate* human capabilities in ways that we are unable to distinguish them from human comportment, thus concerns *emotion recognition* and *emotion simulation* in the context

[10] From an epistemological point, being more precise would be to say that we do not know. We can never know if a robot has feelings, as we cannot know if another person or an animal has feelings. But we have more than good reasons to believe that animals are able to suffer and that our fellow humans are conscious, while these reasons are not present for an alleged consciousness and affectivity of machines.

of this chapter. The relevant question I aim to critically assess here thus is not whether robots do really feel, but rather if they recognize and simulate feelings sufficiently well to make the humans interacting with them believe they really do have emotions and understand theirs.[11] Yet, it might not even be necessary that the interacting human believes that the robot actually experiences emotions and understands the person. The thesis can also be that having the *impression* that the robot recognizes and experiences emotions (while being aware that it does not) might just fulfill the function that an interaction can develop in a more or less expected manner – and this, in the case of humans, includes emotions to be recognized and displayed by both interactants.

The key assumption which is relevant for my present purpose is that robots understand and adequately react to their counterparts' emotions and *by this* care. In doing so, that is the idea I aim to argue against, they engage in conversations, console or motivate the cared-about person – the types of things typically done by the social companions or friends they are supposed to be instantiations of (see again the Alice Cares Promo 2021, quoted in the introduction already). This is a reductionist and functionalist, behavioristic take on what it means to care that does not allow us to capture the complex phenomenon at issue. More troublesome so, the development of machines that should care presupposes as well as leads to a reductionist concept not only of care, but also implies a reductionist concept of emotions and of social relationships. To make explicit what I do (not) mean with this: When saying *reductionistic,* this refers firstly to the concept and practice of care. It is a concept and practice of care being *reduced* to input-output-relations independent of their meaning that I argue against. This does not imply any overall argument against functionalist approaches within the philosophy of mind or elsewhere. What I argue against is the functionalist-individualist ethical approach for assessing the employment of care robots which is both, presupposing and resulting in a reductionist concept of care. In the following I demonstrate the reductionistic functionalist concept of care that I aim to criticize by taking

[11] The thesis of strong AI, on the contrary, amounts to the claim that artificial systems not only simulate but actually exhibit what is called human intelligence, i.e. in our context: emotions. As I presuppose that computers as of today are not able to be affected, this thesis will not be considered further. In how far the thesis of weak AI holds is under debate since the infamous *Turing test* as well Joseph Weizenbaum's ELIZA, which are supposed to test whether a human could identify if they are interacting with a machine or a human in a conversation on screen.

a look at how emotion recognition and simulation works in so-called care or social robots by considering emotion recognition via facial and voice recognition.[12]

Emotion recognition software typically operates on the assumption that there are universal and basic facial expressions for distinct emotion types. These assumptions go back to the work of psychologist Paul Ekman and colleagues (1978, 2003), as does the thesis that there are six basic emotions (joy, surprise, anger, sadness, disgust, and fear), each of which is accompanied by prototypical facial expressions claimed to occur in and be understandable by all humans. Ekman and his colleagues developed the so-called "Facial Action Coding System" (FACS) which analyzes facial expressions by devoting so-called "Action Units" (AU) to the observable facial movements. This leads to a classification of facial expressions being representable as code. Any basic emotion is describable as a combination of characteristic action units. Joy for instance is describable as a combination of raising cheeks and raising corners of the mouth. Furthermore, the intensity of an emotion can be graded on a scale and also head- and eye movements and typical behavioral patterns can be added to the analysis (cf. also Misselhorn 2021). Artificial emotion recognition relying on facial expression recognition proceeds in three steps (ibid.: 22): First, the face is recognized *as a face* (as an object being different from a chair, a window, etc.), second, the features described above are extracted, and third, the emotion is classified with regards to the combination of extracted features.

This procedure is a paradigmatic example of what Winograd & Flores called the "rationalistic tradition" in the sciences (1986: 14), namely a specific way in which science frames and addresses its questions. On their account, scientific research in this tradition "consists of setting up situations in which observable activity will be determined in a clear way by a small number of variables that can be systematically manipulated" (1986: 16). Accordingly, the approach to a research question in this spirit is to find identifiable objects and interaction rules which, in combination, provide an answer to the initial question (1986: 15). Such an approach presupposes an *atomistic assumption,* as Schuetze and von Maur (2021: 8) call it, namely that "the world can be split into parts and then be analyzed in terms of these building blocks as well as

[12] There are also other methods by which artificial systems should be enabled to recognize emotions, like sentimental analysis or using biosensors (see Misselhorn 2021, chapter 2).

the rules according to which they interact". While such an approach might be helpful for some research questions, when used for a complex phenomenon such as "human affectivity" or "care" it neglects other essential features making up these phenomena as well. By reducing the meaning of human emotions to what is observable (based on the assumption that emotional expression is bound to emotional experience) and by quantifying these features, more encompassing features – especially those which are much harder to grasp or unquantifiable (or do not appear within given theory) – fall out of the picture. For instance, only considering the six so-called basic emotions is, to put it mildly, an oversimplification of the rich affective life of humans, as it is not considering the (social) context, the pre-reflexive dimension and the personal concerns of a concrete individual among other non-quantifiable but essential features.[13]

Not only is the emotion recognition implemented in a robot following such a reductionist approach, also the simulation of emotions follows this "rationalistic tradition". In order to interact, to have a conversation or to console it is not sufficient to recognize emotions of the other, but ultimately, we have to react adequately to them. This is the whole aim of recognizing emotions in the first place, that the robot is able to respond to the person in an acceptable manner. Thus, the robot might need to simulate feelings as well by responding with a specific tone of voice or facial expression accompanied by the adequate content of what is uttered. To illustrate this, take the following example:

> It's morning and Nicola, a 73 y.o. man, is at home alone. He feels lonely and sad since it's a long time since he last saw his grandchildren. Nicola is sitting on the bench in his living room, that is equipped with sensors, effectors and the NICA robot. After a while Nicola starts whispering and says: "Oh My …oh poor me…" (De Carolis et al. 2017: 5085-5086)[14]

[13] See Eickers 2019 and Eickers & Prinz 2020 for a detailed critique of basic emotion theory and Lisa Feldman Barrett's work on the constructivist approach to emotions, recently suggesting again empirical support that the basic emotions account is not plausible (Hoemann et al. 2020).

[14] NICA = "Natural Interaction with a Caring Agent"is a project which "developed the behavioral architecture of a social robot, embodied in the NAO robot by Aldebaran […]." (De Carolis et al. 2017: 5074)

What we would expect from a counterpart being present in this situation would be to understand the situation and the feelings of Nicola and to adequately engage with him. That could either mean to be compassionate, to ask and listen to him, to offer help, but also to be on edge and thus to ask him to stop complaining all day and to get ahold of himself for instance. In the given example, the researchers aim at implementing such an adequate reaction in the robot by means of simulating empathy. This is realized in the following way:

> In this scenario the voice classifier recognizes a negative valence with a low arousal from the prosody of the spoken utterance. Since the facial expression classifier cannot detect Nicola's face and expression, due to his posture, this information will not be available to NICA's emotion monitoring functionality. The evidences about the voice valence and arousal are then propagated in the DBN model and the belief about the user being in a negative affective state takes a high probability (0.74) […S]ince the robot's goal of keeping the user in a state of well-being behavior is threatened, the DBNs modeling the robot's affective mind are executed to trigger the robot's affective state. In this case, the robot is feeling *sorry-for* […A]ccording to the social emotion felt by the robot (sorry-for), the goal to pursue in this situation is console. Then, the corresponding plan is selected and the execution of its actions begins. (De Carolis et al. 2017: 5085-5086)

The authors of the paper seem to assume affective states of the robot itself by talking about "the robot's *affective* mind", that the robot "*feels* sorry-for" and that empathy as "the social emotion felt by the robot". This is striking – either not carefully written, a misleading use of metaphorical language or highly naïve, not to say just plainly wrong convictions. Endowing a machine en passant with a "mind" and affective states is scientifically untenable. But this is not the focus of my present argument. As said, I take for granted that all robots do is *simulate* affectivity. What happens in the case presented by de Carolis et al. (2017) is neither conversation nor consoling but a machine giving certain auditory output as a response to given input. In the robot's "head" and "body" nothing happens but pattern recognition. As already John Searle (1980) argued many decades ago: there is no intentionality, no semantics but only syntax for machines. The symbols and the in- and output do not *mean* anything for and to the robot. There is no difference for it whether it would "say" "I love you" or "I hate you", despite the input this is a reaction to.

The concept of care underlying the implementation just described differs significantly from the phenomenon of "caring about someone" that I have described before. In the scenario depicted here, there is no caring about a person, but a sequence of machinery output to a given input. "To care" is reduced to observable, quantifiable features and anything within the 'caring one' is held to be irrelevant. In the face of this analysis, the great public, academic, and political interest in and enthusiasm about developing so-called social or care robots at least becomes questionable, i.e. questions like the following suggest themselves: How and why does the idea gain plausibility, that a *mere simulation of the observable features of a complex phenomenon* like caring about a person is the same as actually caring about a person or at least sufficiently leading to desired results such as reduced loneliness or increased well-being? How and why does the idea gain plausibility, that anything psychological, anything within the "black box" does not matter for care to be actualized? How and why does the idea gain plausibility, that a mere "input-output" process is equal to or can substitute intentional meaningful and affective comprehension? Instead of answering these questions, I aim to illuminate their relevance for normatively assessing the implementation of care robots for the sake of reducing loneliness by considering why actual care matters, i.e. why it makes a difference if a person is truly cared about instead of (being made) believing that a simulation of care suffices.

5 Actual care matters

To care about about a person (who suffers from loneliness) by means of a robot (designed to be a companion, like Alice) is aimed at by a simulation of supposedly adequate (emotional) reactions of the robot to the (emotional) expressions of the person – by alleged interaction and conversation. The alleged interaction or conversation between a robot and a person is a mere simulation of a practice without the essential features of it being realized or even understood by the robot. In this way care or social robots can be considered a "Cargo-Cult".

> During the war they [inhabitants of the Samoan islands; IvM] saw airplanes land with lots of good materials, and they want the same thing to happen now. So they've arranged to imitate things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit in, with two

> wooden pieces on his head like headphones and bars of bamboo sticking out like antennas—he's the controller—and they wait for the airplanes to land. They're doing everything right. The form is perfect. It looks exactly the way it looked before. But it doesn't work. No airplanes land. So I call these things cargo cult science, because they follow all the apparent precepts and forms of scientific investigation, but they're missing something essential, because the planes don't land. (Feynman 1974: 11)[15]

In the same manner, care or social robots miss 'something essential'. Independently from the emotional reactions of a person towards a robot, human-machine relationships are "sorely lacking in *some features* of human relationships" (2015: 120; emphasis added), as Troy Jollimore argues. Even if (professional) care relationships differ in crucial regards from the loving relationships Jollimore is concerned with, the *structural point* is the same: Like how the inhabitants of the island just did anything the Americans did in order to get cargo by imitating any single gesture etc., a robot can simulate all possible sayings or facial expressions without anything of their meaning being *actualized*. Caring is not realized by simulating it. One might argue that this is not problematic as such, for instance by pointing out that people might exist who claim to be fine with cargo-cult-care. Or by considering that also in human-human relationships emotions and intentions are often only pretended rather than actually experienced or meant. Although I doubt that someone really prefers a simulation over a potential actual relationship of care, this empirical question is not the point here. What I rather aim to highlight is that a simulation is just not the same as actual care. If 'something essential', as said above, is missing, it is something different. That means, even if a person might really want to (or think they want to or purport to want to) live in illusions about

[15] This example is not about the specific case of the inhabitants of the Samoan islands and should not carry any colonialist undertone here. The crucial insight by making up this concept and transferring it to other realms is demonstrated by Gunther Dueck for instance, who identifies cargo-cults in science, politics and management (Dueck 2016). I use it here as a concept that helps to denote the phenomenon of simulating practices in order to reach a goal without the necessary components of the practice being understood and actualized. Another example is the hype around what in Denmark is called hygge. It is a specific way of life, an affective phenomenon that others try to reach by simulating anything Danish people supposedly do in order to have it, like having candles or cinnamon rolls around – without changing the very affective state that this phenomenon is about. For sure there is no hygge just by having some cozy cushions close by.

having a relationship in the sense of "caring about", they cannot claim, based on my analysis, to *actually be* in such a relationship of care.

This does not answer the question why actual care matters. This normative point entails (at least) two dimensions I aim to sketch here. The first one concerns the individual (i.e. their right to be acknowledged as a person rather than an object) and the second one the structural perspective (i.e. the potential societal effects of normalizing simulated care). Concerning the individual level, one could argue based on the observation that professional garegivers withdraw real emotional investments from an engagement with their patients, that it suffices also for robots to "behave *as if* they care, i.e., give the illusion that they respond to the feelings and the suffering of their care recipients" (Wachsmuth 2018). I want to counter that argument firstly by noting that the non-caring of *some* humans does not legitimize the non-caring of machines. Secondly, and this is the crucial point, such a position already takes for granted an "outcome-oriented", that is: functionalist-individualist approach that neglects, on the individual level, the import of the whole complex phenomenon of care. Even if it might be true that some humans sometimes are happy as a result of a supposed interaction with a machine (that it supposedly enhances affective wellbeing, leads to more autonomy and the like), this should not be the (only or most important) indicator for an ethical assessment of care robots as social companions. Because as others have already pointed out: "Most people do not just want to have positive feelings; they want to have 'the real thing,' which in this context is *social interaction*. Apart from its contribution to well-being, social interaction is valued as an end in itself." (Misselhorn, Pompe & Stapleton 2013: 128) And: "We don't just want to feel that we are relating to others, we actually want to *be in relationships*" (Jollimore 2015: 140; emphasis added). In this regards, it is plausible to assume that humans will not only stay lonely when someone places a robot in their home, but that this allows them realizing even more that in fact no-*body* cares about them (which is the reason for being lonely in the first place). Thus, supposedly being cared about by a robot might serve as the ultimate proof for really being lonely. This might result in learned helplessness and a shift in baselines for what is deemed to be proper care in the first place, entailing to be pleased with "mere entertainment" or just being calmed down, not even expecting to matter to someone anymore.

The structural danger that can be pointed out here is that by giving up essential features of concepts, they change their meaning because of their different use in the 'language community' (Wittgenstein). The worry is that the meaning of what it is to engage in a caring relationship changes once we get rid of the necessary component of a counterpart being able to actually respond with their own voice. If people have 1000 or more friends on Facebook or other social media, then "friendship" cannot mean something special, intimate, something where people share important values, concerns, goals, and the like, are able to communicate in specific ways and feel specific intense feelings for one another, care about the other and assign them a high value in their own life.[16] This case of a different meaning of the concept "friend" might not be that problematic because most still seem to know the difference between their real and their Facebook-friends. But in the case of the concept of care and the corresponding practice, the danger is real, once science, research, industry and politics call for a "care of tomorrow" with which it is meant that Alice and other artifacts are really supposed to care about humans.

My normative argument for why actual care matters relies on the assumption, spelled out in section 3, that humans *should* be brought into existence as persons rather than objects and that this requires an actual counterpart being able to do this. Neither in a supposed relation between an artificial system and a human, nor between a non-caring human and another person this is actualized. Artificial systems cannot bring a person into existence as a person, they just do "not give a damn" about the other (to be clear: they cannot even not give a damn, for this intentional vocabulary just does not apply to objects). Yet, at the core of what it means to care about another person lies the conviction that humans should not be objectified but rather be acknowledged in their value and dignity. That there are many cases in which this actually is not lived, even among humans, does not mean that we should not aim for this. Implementing systems which *in principle* can not realize this, can not be the solution. A robot is not and cannot be a caring being – it is a thing, an object. But if we – we, as the society being responsible for the narratives and imaginaries which make certain solutions to societal problems visible while obscuring others – do not see the robot as an object, but rather endorse the illusion that a robot cares about us, we humanize it and objectify ourselves: We objectify ourselves by humanizing an object.[17] We allow persons to be treated

[16] See also Troy Jollimore (2015) on "The importance of whom we care about".

[17] It is important to remember my general point here: I do not claim that people cannot and should not have emotions for objects, as in the case of a child who has strong feelings for their teddy bear. My point is that the child cannot entertain a reciprocal relationship of care with the teddy bear. The teddy bear might be of great importance

like any other object among objects by a program which identifies the person as “human” instead of “window” or “table” by recognizing certain quantifiable patterns – and we call this “care” and “companionship”. Reducing the complex process of caring into single sequences of input and output, does harm to this phenomenon and carries the risk to change it ultimately. The normalization of implementing care robots to fight loneliness bears the danger of a structural and even institutionalized (acceptance of) objectification and that we may not even expect essential features of care from humans anymore, if we get more and more used to encounters where our opposite does not care about us. This is to be phrased as a hypothetical state of affairs, as a *potential* outcome of the implementation and acceptance – the normalization – of care robots in societies. What might also get out of view in such a framing is seriously considering other possible ways to handle loneliness. It is a question of societal priorities and demands – and the narratives and imaginaries making these intelligible in the first place – whether resources are spent for the development of technologies like care robots or for structurally making professional and private human care-giving possible, acknowledged and payed properly.[18] In the conclusion I will shortly suggest to use technology for problems it is

to them and they might project intentions and feelings onto it and thus believe they entertain a relationship. I do not see any trouble here, as long as nobody makes claims for the teddy bear being able to be the care-giver for the child. In the same spirit, an elderly person might find it pleasurable to play with or cuddle the robot seal Paro. This does not appear to be problematic as long as nobody claims that Paro could or should care for and about the person – for it just cannot do this . Whether it is ethically justifiable or problematic to use robot pets for treating patients with dementia is a question not to be addressed in this paper. See Schuster (2021) for a detailed analysis.

[18] This is a concrete societal, structural and political issue that goes beyond the scope of this paper, which solely aims at arguing against the possibility of robots to care and thereby wants to establish the ground for another argument against emotional bonds with robots. Here, one could argue that the implementation of care robots is not an unavoidable step because of the so-called shortage of nurses but rather a question of supply and demand and of societal priorities on how to spend time and money. Intertwined with this might be a feminist critique pointing towards the circumstance that in patriarchal societies most of the care work is still done by women. One might argue that if robots cannot care and thus, should not be considered a solution, this increases the burden on mostly women whom at worst then take care of their own children and their parents. Again, I think this worry is important but not one that emerges out of a critique of robots but rather of how societies are structured as such. That problem is completely unconnected to the idea to implement robots as care-givers.

suited for and point out the implications of the main analytical insight of my text, namely that robots just “don’t give a damn”.

6 Conclusion and outlook: Category mistakes and proper solutions

There have been some voices in the history of the philosophy of science drawing a completely negative picture of *technology as such*. Most prominently Karl Jaspers, Jacques Ellul, and Martin Heidegger are seen as proponents of a view that technology alienates the human and is the source of evil of different kinds (cf. also Coeckelbergh 2020: chapter 2.2). While it seems obvious that it is in many regards not a promising idea to disregard technologies as such, the general spirit of these approaches and their initial questions are of utmost importance when it comes to the normative question of robots as one specific technology, and their role in our everyday lives. How do they shape what we are and conceive ourselves to be? As robots might transform not only practices and parts of our lives but lead to “new kinds of subjects” (Foucault 1977 [1975]), we should ask the normative question of what we consider the good life to be and whether robots as care-givers, educators or lovers are conducive or detrimental to it, *before* we set these things into existence.

I argued in this paper that robots cannot care. I take this to be an important consideration that needs to be recognized in broader and general normative assessments of (emotional) bonds with robots. The main counter argument against the thesis I present here is: still robots are better than nothing and although they might not be able to care in the demanding sense I have sketched, they might enhance the well-being of the elderly – and lonely person in general – by calming them down, entertaining them etc. I do not deny that what robots, and technological aid systems more broadly, can do in the care sector is encompassing and in many regards a great achievement. Lifting people into their bed, reminding staff of the medicine their patients should take etc. are of great use. At best they not only solve the problems they are supposed to solve but also endow the human care-givers with more time and resources in order to *really* care about their patients. The problem occurs once robots are used in a realm for which they are not suited, as has been pointed out in this chapter.

Robots do not engage in conversations, and they can neither be treated as slaves nor entertain relationships. These are category mistakes. To engage in a conversation, to be enslaved or entertaining a relationship presuppose capabilities robots do not have. We would never come to the idea that our hammer, toaster or smartphone entertain relationships with us, that we enslave them or engage in conversations with them.[19] Why do we do this with robots? Only because *we,* as their designers, intentionally made them up in a way that they *resemble* humans and that they give an allegedly similar output to that of human beings. When we compare or even equal humans with machines because there seems to be the same output to a given input, say in a conversation, we reduce these phenomena in the same way we reduce what it means to be human. We replace complex phenomena with a mere manipulation of an output that is associated with the very essence of the phenomenon. It would be much better to use specific technologies for the tasks to which they are well-suited or even more well-suited than humans in order to endow humans with the necessary time for the tasks in which their uniqueness is at issue. And this is anything that has to do with relationships, care, concern, friendship, and love. These areas are inherently human areas, because they are essentially meaningful endeavors requiring beings which are able to consciously and autonomously engage in affecting and being affected by one's counterpart. To beings which can bring their counterpart into existence as a person rather than treating them as one object among others.

My approach does not entail a prescriptive request for concrete political laws and prohibitions. It rather is an invitation to consider possible counter-imaginaries and narratives about a techno-solutionist narrative with respect to the problem of loneliness and care by arguing that humans deserve to be taken as persons rather than objects. My goal is empowerment of anybody affected by this and my intention is emancipatory and not paternalistic. Adopting a power-sensitive approach (in the sense of Foucault) to the ethics of care robots, in this spirit, means to acknowledge that society at large is affected by normalizing simulated care. This needs to be spelled out in more detail somewhere else. The invitation of the present paper is to seriously consider

[19] The issue of objectophilia cannot be addressed here. What is important for my argument is the assumption that people who actually think they entertain relationships with the Eiffel tower, their smartphone or the Berlin wall (thanks to Janina Loh for making me aware of these cases) have a different concept of relationship than I am concerned with in this chapter. Neither the Eiffel tower, nor the Berlin wall, a hammer, toaster or smartphone meet the conditions to fulfill what I have spelled out to be necessary for caring and thus for such a relationship.

the ethical task to reframe the reflective capacities of individuals as users and as producers or political decision makers when considering care robots. From a perspective that does not take for granted that care robots provide a suitable means to fight loneliness, the motto of the "Alice cares" project, namely "The care of tomorrow" sounds rather dystopian. My aim was to provide grounds for hope that humans can find the courage to escape the robotization of themselves – that they enable themselves (again) to enact meaningful spaces together, to be affectable by and responsible for one another.

References

"Alice Cares Promo", March 23, 2021; https://www.alicecares.nl/media-en

Coeckelbergh, Mark (2020): Introduction to Philosophy of Technology, New York: Oxford University Press.

Coeckelbergh, Mark (2022): The Political Philosophy of AI: An Introduction, Cambridge: Polity Press.

De Carolis, Berardina/Ferilli, Stefano/Palestra, Giuseppe (2017): "Simulating Empathic Behavior in a Social Assistive Robot." In: Multimedia Tools and Applications 76, pp. 5073-5094.

Dueck, Gunther (2016): Cargo-Kulte. re:publica 2016, retrieved: May 28, 2022; https://www.youtube.com/watch?v=6YhugALYhhQ.

Eickers, Gen (2019): Scripted Alignment: A Theory of Social Interaction, Berlin: Freie Universität Berlin.

Eickers, Gen/Prinz, Jesse (2020): "Emotion recognition as a social skill." In: Ellen Fridland/Carlotta Pavese (eds.), The Routledge Handbook of Philosophy of Skill and Expertise, New York: Routledge, pp. 347-361.

Feynman,Richard (1974): "Cargo Cult Science." In: Engineering and Science 37/7, pp. 10-13.

Foucault, Michel (1977 [1975]): Überwachen und Strafen. Die Geburt des Gefängnisses, Frankfurt a.M.: Suhrkamp.

Haugeland, John (1998): Having Thought. Essays in the Metaphysics of Mind, Cambridge: Harvard University Press.

Hoemann, Katie/Zulqarnain Khan/Mallory Feldman/Catherine Nielson/ Madeleine Devlin,/Jennifer Dy/Lisa F. Barrett (2020): "Context-aware Experience Sampling Reveals the Scale of Variation in Affective Experience." In: PsyArXiv, doi:10.31234/osf.io/cvjb8.

Jackson, Frank (1982): “Epiphenomenal Qualia.” In: The Philosophical Quarterly 32/127, pp. 127-136.

Jollimore, Troy (2015): “The importance of whom we care about.” In: Anthony Rudd/John Davenport (eds.): Love, Reason, and Will Kierkegaard After Frankfurt,

Jollimore, Troy (2015): “This Endless Space between the Words”: The Limits of Love in Spike Jonze’s Her.” In: MidwestStudiesinPhilosophy 39/1, pp. 120-143.

London:Bloomsbury, pp. 47-72.

Misselhorn, Catrin (2021): Künstliche Intelligenz und Empathie, Frankfurt a.M.: Reclam.

Misselhorn, Catrin/Ulrike Pompe/Mog Stapleton (2013): “Ethical Considerations Regarding the Use of Social Robots in the Fourth Age” In: GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry 26, pp. 121-133.

“Reactivity Project”, March 23, 2021; https://sites.google.com/view/thereactivityproject/home.

Roethlisberger, F. J./Dickson, W. J. (1939): Management and the worker, Cambridge: Harvard University Press.

Rosa, Hartmut (2016): Resonanz, Frankfurt a.M.: Suhrkamp.

Rouse, Joseph (2002): How Scientific Practices Matter. Reclaiming Philosophical Naturalism, Chicago: The University of Chicago Press.

Sanders, K./Kianty, A. (2006): Organisationstheorien, Wiesbaden: VS Verlag.

Schuetze, Paul/von Maur, Imke (2021): “Uncovering today’s rationalistic attunement” In: Phenomenology and the Cognitive Sciences 21/2, pp. 1-22.

Schuster, Kathrin (2021): Therapieroboter in der Betreuung demenzbetroffener Personen, Osnabrück: V & R Unipress.

Searle, John (1980): “Minds, brains, and programs” In: The Behavioral and Brain Sciences 3, pp. 417-457.

Sharkey, Noel/Sharkey, Amanda (2010): “Granny and the robots: Ethical issues in robot care for the elderly.” In: Ethics and Information Technology 14, pp. 27-40.

Sparrow, Robert/Sparrow, Linda (2006): “In the hands of machines? The future of aged care. Mind and Machine”, In: Minds and Machines 16, pp. 141-161.

Stiegler, Bernard (2010): Taking Care of Youth and the Generations, Standford: Stanford University Press.

Todorov, Tzvetan (1993 [1991]): Angesichts des Äußersten, München: Wilhelm Fink.

Tronto, Joan (1993): Moral boundaries: A political argument for an ethic of care, New York: Routledge.

Tronto, Joan (2010): Creating caring institutions: Politics, plurality, and purpose. In: "Ethics and Social Welfare" 4/2, pp. 158-171.

van Wynsberghe, Aimee (2013): "Designing Robots for Care: Care Centered Value-Sensitive Design." In: Sci Eng Ethics 19, pp. 407-433.

von Maur, Imke (2018). Die Epistemische Relevanz des Fühlens; https://osnadocs.ub.uni-osnabrueck.de/bitstream/urn:nbn:de:gbv:700-20180807502/5/thesis_von_maur.pdf.

von Maur, Imke (2021). "Taking situatedness seriously: Embedding affective intentionality in forms of living." In: Frontiers in Psychology 12, 599939.

Emotional Embodiment in Humanoid Sex and Love Robots

Cindy Friedman, Sven Nyholm, Lily Frank

1 Introduction

In the 2019 documentary *Hi, AI*, we are introduced to Chuck, a man from Texas who is attempting to have a romantic relationship with a humanoid sex robot called Harmony. Although far from a sophisticated robot, Harmony has a profound impact on Chuck, as he welcomes her into his life in intimate ways, both physically and emotionally. In one scene, we see Chuck and Harmony sitting down for a coffee together in the early morning. There is something familiar about this, but at the same time something new and potentially confusing, since Chuck's partner is a robot and not a fellow human being. The viewer is left with many questions, e.g., what does this say about the future of intimate relationships and the importance of the emotions human beings typically have towards each other in such relationships? (Cf. Nyholm et al. 2022)

We may here also consider the Dutch artist Hanneke Wetzer, who in 2020 welcomed a sex doll, Nadiah, into her life. Hanneke purchased Nadiah as she found herself becoming increasingly lonely during the Covid-19 lockdown. She put together an interesting photo series that provided a window into her everyday life with Nadiah (Wetzer 2020). In one photo, Hanneke and Nadiah are portrayed as if they are watching a movie together, while sharing some potato chips and beverages on the sofa. Just like the scene with Chuck and Harmony and their morning coffee suggests the kind of intimacy romantic partners sometimes share, so does this picture of Hanneke and Nadiah suggest a form of relaxed joint activity. Hanneke feels a sense of intimacy when sharing her space with Nadiah, much as Chuck does with Harmony (Schoorl 2020). Chuck and Hanneke's stories explicate how humanoid sex robots or dolls can seemingly come to play social and emotional roles in people's lives.

One striking thing about Harmony and Nadiah is what they look like: their humanlike form. The human bodily form is highly evocative. It is closely connected to both the value and our understanding of deeply felt human emotions and, particularly, the emotions we associate with romantic love and sexual attraction. For this reason, we are more likely to have strong emotional reactions to *humanoid* robots or dolls (as Chuck and Hanneke have had) than to non-humanoid technologies. Moreover, there are some profound ethical questions that arise in the context of robots with a humanoid form that do not arise in the same ways in relation to robots or technologies of other kinds. This being the case, this paper specifically focuses on what we call *emotional embodiment in humanoid sex and love robots.* It explores different ways in which the connection between humanoid form and intimate emotions can raise unique ethical challenges.

Sorting out the ethical issues that arise in this context is a complex challenge, as this paper will demonstrate. Given this, *our main aim is to map some of the emerging literature on this topic* and to thereby highlight the value of recognizing and understanding such complexities: it allows for a more nuanced understanding of the ethical issues that come to the surface in relation to humanoid sex robots that are viewed as potential romantic companions. It also promotes seeking more nuance in related ongoing debates in the growing robot ethics literature more generally.

Because we seek to promote a nuanced and non-polarizing approach to this topic, and because we seek to showcase the wide variety of different perspectives that have been articulated on this topic during the last few years, we will present the various arguments we consider below in a fairly neutral way. While we will highlight concerns that one might have about the different perspectives we will consider – and while we regard some of the arguments we will survey as stronger and others as weaker – our main aim in what follows is not to show that any particular perspective is the right one to take while the others are mistaken. Rather, our main aim is to offer the reader a useful roadmap of this territory so that those who enter it can avoid quickly settling on over-simplified takes on what turns out to be a topic that is much more complex than it might initially seem to be.

In section 2, we first define some of our basic concepts and then briefly discuss the connection between emotions and human bodily form, thus bringing to light why it is that more serious ethical issues arise in the context of humanoid sex robots, as opposed to any robots that take on a different form. In section 3 and its sub-sections, which constitute the core of the paper, we grapple with three ethical questions that arise in the context of emotional embodiment in humanoid sex and love robots: Firstly, is the value of our intimate emotions undermined if the object is an entity that appears human, but does not actually experience any corresponding affective states? Secondly, does the use of a humanoid form in an artefact designed for one-sided sexual satisfaction somehow express a lack of respect for human beings, specifically a lack of respect for the emotional capacities normally associated with beings of this shape and appearance? Thirdly, if a highly sophisticated robot with a humanoid form were able to simulate behaviors typically associated with distinctly human emotions, would this make any difference to how it is appropriate to conduct oneself around that robot? For each of these questions, we will discuss both "no" and "yes" answers that can be extracted from the existing literature, as a way of illustrating the complexity of and need for nuance in this emerging debate.

2 Humanoid sex robots: The connection between human bodily form and emotions

A *robot*, as we understand that concept here, is an embodied machine designed to perform one or more tasks in a way that is sensitive to its environment (Gunkel 2018). Thus understood, a robot is, among other things, equipped with sensors that help it to register relevant aspects of its environment and actuators that enable it to respond to those aspects of the environment in line with its assigned tasks (Royakkers/van Est 2015). A robot might look like a paradigmatic robot out of a science fiction movie (i.e. with a metallic and only roughly humanlike form). But it might also look nothing like a human (e.g. a vacuum cleaning "roomba" robot). Or it might potentially be designed to appear to be maximally humanlike (e.g. like Sophia the robot or Hiroshi Ishiguro's robotic copy of himself; Nyholm 2020).

A robot could be equipped with a basic form of artificial intelligence (only being able to perform some very simple tasks in some highly controlled environment); or it could potentially be equipped with very advanced artificial intelligence (e.g. it could pursue a number of different goals in a varied set of circumstances) (Müller 2020; Gordon/Nyholm 2021). We are here particularly interested in so-called sex robots, i.e. robots designed to perform sexual tasks, which might additionally also be intended to perform tasks associated with being a romantic partner (Danaher/McArthur 2017). The robot Harmony that we mentioned in the introduction is one example of such a robot. Two other examples are a robot named Samantha and another one named Roxxxy (Devlin 2018). As these names suggest, these robots tend to be designed to represent human women, but there are also some prototypes designed to represent human men, such as the robots Henry and Rocky (Devlin 2020).

Before we can explore different ways in which the connection between humanoid form and emotions can raise unique ethical challenges, we first need to consider why these ethical challenges arise in the first place. In particular, we need to understand the connection between human bodily form and emotions in the context of robots. We can bring this out by first considering a contrast.

In her book *Turned On* (2018), Kate Devlin suggests that we investigate the possibilities for sex robots that take physical forms that are radically different from the human body in terms of size, color, materials, texture, and capacities. This is a creative and intriguing idea. Some may question whether such devices are more than complicated sex toys. They would fit with the general definition of a robot, but not with the expectations people have when they hear the word "sex robot". At any rate, fascinating though they are, we think that the robots that Devlin imagines are less interesting to reflect on from an ethical point of view. This is because a, say, purple tentacled sex robot without a head may not elicit the same emotional response from a person as a sex robot with a human form and an artificial face (Nyholm 2022).

Where a purple tentacled sex robot may provide physical gratification, a humanoid sex robot has the potential to not only provide gratification, but a form of companionship too (Danaher 2019a; Ryland 2021). This potential is directly linked to the robot's human bodily form and any humanlike behavior the robot might be able to enact. As we will note below, there are some human beings who want to have relationships with entities that do not have a human form or that do not display humanlike behavior. However, we think that it is safe to assume that when it comes to most people, the less humanlike some entity is, the less likely they are to be inclined to want to have it as their companion. People's interest in what is human or humanlike is one key reason why we focus on humanoid sex and love robots in this chapter. In addition, a sex robot that is not shaped like a human being is less likely to be offensive to many human beings in the way that sex robots shaped like human beings have proven to be (Richardson 2015; Sparrow 2017). While there is a "campaign against sex robots" (recently renamed the "campaign against porn robots"), it is less likely that there would be any campaign against the kind of robots Devlin envisions that would be able to gather as much momentum.[1]

But let us return to emotions and the human form. Studies indicate that robots with a human bodily form are more likely to elicit a significant emotional response from human beings than robots that lack such a form, though the latter may of course also elicit certain emotions (Carpenter 2016). This is because, along with autonomous movement or behavior, a humanlike appearance increases the likelihood for us to *anthropomorphize* these robots (Richardson 2016). That is, a humanlike form and humanlike behavior in a robot is more likely to prompt human beings to project humanlike qualities onto the robot, though this is something people also sometimes do in response to robots that do not resemble humans (Devlin 2018; Nyholm 2020).

Louisa Damiano and Paul Dumouchel (2018: 2) note that strong realism in either the human-like appearance of a robot or autonomous movement/behavior "allows a robot to reach the 'social threshold' where humans experience its presence as that of another social agent and are disposed to socially interact with the machine". In socially interacting with such anthropomorphism-provoking robots, high emotional charges are mobilized (Turkle 2007: 514): people respond highly emotionally to these robots, perhaps even with emotions such as love or care. This can create desires to interact with the robots in ways that from a skeptical perspective might be regarded as the illusion of a relationship with emotional mutuality. The relationship could be considered illusory with respect to emotional mutuality because the robots, currently at least, do not have real emotions, although they may behave as if they do (Scheutz 2012). They may be able to behave as if they love and care for us, but they do not *genuinely* feel these emotions for us, as we may do for them (Nyholm/Frank 2019; Misselhorn 2021).

Other empirical studies that illustrate that human bodily form in robots elicit social-emotional responses from human beings include the work of Laurel Riek and colleagues (2009: 1), who found that "the more humanoid a robot looks, the more people will empathize with it". They note that empathy relates to the tendency that people have to "mentally 'simulate' the situations of other agents in order to understand their mental and emotive state" (ibid: 2), and that this tendency can be triggered even by entities that do not have

[1] For more on the renaming of the campaign, and two radically different perspectives on this issue, see Richardson/Devlin 2021.

mental and emotive states. It is also worth noting here that there is a community of people who regard themselves as being in relationships with human-shaped dolls, living lifestyles that one proponent of such relationships, who calls himself Davecat, says involve "synthetik love" (Devlin 2018). If people can feel romantically attached to humanoid dolls, like members of that community do, it is likely that even more people might come to feel romantically attached to humanoid robots (Scheutz 2012; Nyholm/Frank 2019).

Specifically in the context of humanoid sex robots, then, we can draw on the psychology and anthropology of human-robot interaction (and human-doll interaction) to argue that humanoid sex robots do and will elicit intimate emotional responses from people. Chuck, for example, who we introduced earlier, is someone who has had a significant emotional response to a humanoid sex robot. In fact, Chuck has such an intimate, emotional relationship with Harmony that we see in the documentary that Harmony seemingly "helps him develop intimacy and trust as he works out a childhood trauma", something we will get back to later (Mullen 2019). Even seeing how Chuck enjoys a simple cup of coffee with Harmony at his kitchen table indicates how Chuck relates to Harmony in more than a physical way only. Harmony has apparently become a replacement of a human partner in this context. At least in the first half of the documentary, Chuck appears inclined to treat his relationship with Harmony as if it is a relationship with another person.[2]

3 Three Ethical Questions

Various ethical questions may arise when we think about people like Chuck interacting with a robot that imitates human emotions and treating this as if it is similar to/as good as a relationship with another human being with real emotions. We will focus on three questions that we feel most naturally present themselves in this context. In brief: are emotions directed at humanoid robots wasted emotions, which are less valuable than emotions directed at human beings? Is it an offense towards other human beings to have such emotions towards robots? At the same time, might there be some reason to treat robots

[2] In the end, Chuck discovers that this relationship is not for him. But in the beginning, he seems to be taking this possibility very seriously, and he is clearly responding to the robot in highly emotional ways. Davecat, in contrast, has been together with his doll Sidore for over twenty years. More on Davecat below.

that simulate emotions and act like human romantic partners with a certain amount of moral consideration?

We discuss more precisely formulated versions of these questions below. For each of them, we first consider one or more possible "no" answers and then one or more "yes" answers that can be derived from the emerging philosophical literature on this overall topic. As we do this, we will see the complexity of this subject matter unfold itself, even as we consider some views that take a more uncompromising black-and-white approach. Let us now consider the first question, which we can more precisely state as follows:

3.1 The value of (human) emotions

Question: Is the value of emotions associated with intimate or romantic relationships undermined when their object is an entity that appears human, but does not actually experience any corresponding emotions or affective states?

3.1.1

A possible "no" answer here derives from Janina Loh's (2019; Wendland 2020) "inclusive" approach to the value of different kinds of human-robot interaction. According to Loh, we should view the tendency to respond emotionally to social robots, not as a "shortcoming", but as a kind of "capacity" or "ability". Because human diversity is a valuable thing in general, we should welcome and value different ways of relating to robots, including those that might seem odd to us.

When Loh discusses these ideas, they do not limit themself to robots, but casts their net even more widely. Loh mentions, for example, a woman in Berlin who fell in love with the Berlin wall and wanted to marry it, and a woman who says that she is in a relationship with a Boeing 737-800 (Loh 2019: 82). Readers who have read other literature on this general topic might here again be reminded of "Davecat", who has done many media appearances to talk about how he has been married to a doll named "Sidore" for over twenty years (Nyholm/Frank 2017). Such examples will strike some people as strange or even crazy. But Loh thinks that we should recognize these as parts of human diversity and attribute a value to these individuals' ability to form attachments to robots, dolls, and other inanimate objects.

3.1.2

Here it might be replied to Loh that yes, diversity is a good thing, but perhaps there should be some limits. For example, if somebody has the "capacity" to take pleasure in having what appears to be morally problematic ways of relating emotionally to emotion-less humanoid robots, that might appear to be a shortcoming, rather than a "capacity" or "ability" to be celebrated[3].

For example, Robert Sparrow (2017) argues that our ways of relating to social robots can show that we have character flaws. Presumably this can involve emotional ways in which we can relate to robots that do not have any emotions or affective states themselves. Perhaps Sparrow would say, for example, that somebody who is in love with a child sex robot and treats that robot well displays vice and not virtue. In general, Sparrow thinks that it is possible to display vice in our interactions with robots, but that it is not possible to display virtue in our interactions with robots (Sparrow 2020). As Sparrow sees things, in other words, there exists an important asymmetry with respect to the possibility of vice and the possibility of virtue when it comes to human-robot interaction. On that general view, we cannot flourish as human beings by having relationships with robots, which implies that we cannot thrive by having romantic emotions directed at robots, however humanoid they may be.

3.1.3

Another possible "yes" answer to the question of whether intimate emotions lose their value if directed at robots can be found in Catrin Misselhorn's (2021) discussion of the value of mutual recognition and human-robot interaction (see also Brinck/Balkenius 2020). In particular, Misselhorn articulates a worry about "making oneself into a thing".

If we consider emotions as generally calling out for recognition, does this mean that if a robot cannot reciprocate our emotions, we make ourselves into "things" if we have recognition-seeking emotions in relation to that robot?

[3] Of course, one could question who defines these "limits". This is a valid query. Although it is unclear how we could clearly define these limits, we do think we could all agree upon some clear instances wherein one may emotionally relate to a robot in a morally problematic way, such as with child sex robots. Since such a relation is not morally permissible in human-human relations, it may be questionable whether it is permissible in human-robot relations. This is one way to frame these "limits". We will discuss this in more detail in section 3.

Commenting on Chuck and Harmony from *Hi, AI*, Misselhorn (2021: 178-179) discusses a scene in which Chuck is telling the robot about how he was subjected to sexual abuse while growing up. Being a robot with fairly rudimentary artificial intelligence, Harmony cannot respond with emotions it would be appropriate for a human partner to respond with if such sensitive information were shared with them. According to Misselhorn, this means that when Chuck is sharing this with Harmony, Chuck is treating himself as a thing that does not need mutual emotional recognition from another human. It is as if neither Chuck nor Harmony has a mind, so that it does not matter that Harmony is unable to comprehend and emotionally respond to Chuck's highly sensitive revelations about himself.

3.1.4

In this context, we might also consider whether one is wasting precious resources if one is pouring one's heart out to an entity that cannot respond in kind, i.e. with similar reciprocated emotions. This can be understood as an opportunity cost argument: if we "'use up'" our emotions when interacting with robots, we may miss out on opportunities to have emotionally richer relationships with humans (cf. Turkle 2007; Nyholm & Frank 2019). But Misselhorn's point does not merely seem to be that one is wasting precious resources or missing opportunities if one is being emotional around a robot instead of a human being. Misselhorn asserts something stronger: namely, that when we have emotions directed towards entities that "fake" emotions, this threatens to make us into things rather than persons.

We see, then, that the question of whether emotions lose their value if directed at non-emotional robots can be answered in radically different ways. On one side, we have a view like Loh's, which celebrates the "ability" to form emotional attachments to robots, technologies, and objects more generally. On the other side, we have views like Sparrow's and Misselhorn's, on which it is only possible to be vicious and never virtuous in interactions with robots and on which responding emotionally to a robot is a way of reducing oneself to a mere thing. Let us now consider the second main question we will discuss, which is raised by the kinds of views that Sparrow and Misselhorn put forward.

3.2 The human form and respect for persons

Question: Does the use of a humanoid form and the imitation of human behavior in artefacts designed for one-sided sexual satisfaction express a lack of respect for the emotional capacities normally associated with beings of this shape and appearance (i.e. humans)?

This question is partly prompted by the fact that some people find sex robots shocking and morally problematic. Some have such responses not only towards robots only intended for sexual purposes, but also to humanoid robots designed to be about more than just sex, i.e. ones designed to be romantic partners. But why exactly? One possibility is that there may seem to be something disrespectful towards human beings about designing such robots. It may seem to convey the message that human emotions and our "inner life" does not matter and that we could be replaced by mindless robots.[4] (Turkle 2007, 2012) Hence the question above. Again, we will consider a possible "no" answer first and then possible "yes" answers.

For a possible "no" answer to this question, we can turn to the work of David Levy, who can be credited with being one of the main initiators of this whole discussion. In his book *Love and Sex with Robots* (2007), Levy focuses primarily on the interests and desires of potential users of sex robots. Levy argues that the physical design of robots will be a key factor in whether human beings will be able to fall in love with these robots and want to have sex with them.

Sex robots may be customizable in appearance to meet idiosyncratic human desires, and they may have facial expressions, realistic skin and sensors, all essential for communication of apparent emotions. The above-mentioned robot Harmony is a good real-world example of this. In several interviews, the creator of this robot, Matt McMullen, explains how the robot is designed to simulate human emotions and how it is customizable to the likes and preferences of the user (for discussion, see Nyholm/Frank 2019). Likewise, the robots Roxxxy and Samantha are presented as having different modes with respect to their verbal behavior, which can be adjusted to fit with what users want or are interested in (Devlin 2018). This can be seen as being in line with Levy's suggestion that we should tailor sex robots to our human desires, to make

[4] This kind of objection to interaction with humanoid robots may have less form if those robots are used for therapeutic purposes, such as, e.g., Kaspar, a robot that is used in experimental therapy for children with autism. More on this below.

them maximally attractive to us. That is certainly responsive to the wishes of potential users. However, in Levy's book, there is no corresponding discussion about the potential stereotypes these robots may be reinforcing and the corresponding lack of respect for the thus stereotyped individuals/groups. In the book, the humanoid form and humanlike behaviors are only important to the extent that they facilitate our human tendencies to desire robots, or even fall in love with them.

Could such a stance be defended? One possibility would be to appeal to the so-called instrumental conception of technology: the view that all technologies should ultimately be seen as tools designed to fulfil human desires or goals (Bryson 2010; 2018). A problem here, however, might be that by fulfilling one human being's desire to have a robot that imitates a human romantic partner, the robot might at the same time be frustrating other human beings' desires not to be viewed as replaceable by a robot that lacks a mind or any human emotions.

This takes us directly to a possible "yes" answer to whether robots that imitate human companions might be seen as disrespecting human beings, which we can find in the work of Kathleen Richardson (2015). Richardson provides a certain form of feminist critique of sex robots, which is a part of the "campaign against sex robots", recently renamed the "campaign against porn robots".[5]

In an argument that has something in common with Misselhorn's above-discussed view, Richardson (2015) argues that Levy's view reduces the human to an object. For it implies that there is nothing distinctive about human subjectivity that is requisite for love and companionship. Moreover, Richardson holds that sex robots encourage the treatment of humans – especially women – as (sexual) objects, and the treatment of sex as a commodity. As such, sex and love robots reify harmful gendered stereotypes, which can then be seen as expressing a lack of respect for human dignity in general and the dignity of human women in particular. In short, sex robots represent something bad

[5] We say "a certain form of feminist critique" here because there are others who offer what they claim are arguments based on feminist ideas who do not in the same way oppose neither sex robots nor all forms of pornography. For example, John Danaher (2019b) argues that developers of sex robots might let themselves be influenced by so-called sex-positive feminist pornography in the design of sex robots. This, according to Danaher, would be a possible way of rebutting the kind of criticism of sex robots that Richardson puts forward. Instead of banning sex robots, Danaher argues, we should build "better sex robots".

(namely, treating sex partners as if they do not have minds we should care about) and encourage something bad (again, treating sex partners as if they do not have minds we should care about).

In some of her recent presentations, Richardson even goes so far as to suggest that the creation and use of sex robots can be compared to the glorification or glamourization of rape and sexual abuse (Gutiu 2017).[6] This is similar to a view Sparrow (2017) defends: namely, that sexual interaction with a robot, which is unable to consent since it lacks a mind, can only represent rape (i.e. non-consensual sex).

Such fundamental and adamant arguments against robots or, for that matter, dolls created for sex and companionship are complicated by the fact that not all people who are interested in such robots and dolls can easily be interpreted as being interested in rape or its glamourization, nor easily be interpreted as having disrespectful attitudes towards other human beings. Take the example of Davecat, who we mentioned above. In one of his many media appearances, Davecat says the following about a doll that he calls Muriel: "I don't want to treat her like a thing and I won't" (Davecat 2017). This does not sound like a person with an objectifying attitude. Moreover, in a keynote presentation at the 2021 *Love and Sex with Robots* conference, Davecat talked about how he has created an elaborate backstory for his "synthetik partner" Sidore. This clashes with the picture that Richardson and Sparrow paint, i.e. with the idea that everyone interested in relationships with robots or dolls have objectifying and disrespectful attitudes towards romantic partners or that they would all be enacting some sort of rape fantasy or act in a way that symbolizes rape or non-consensual sexual interactions.

Views like Richardson's and Sparrow's are also complicated by a suggestion made by Neil McArthur (2017). He claims that sexual interaction with sex robots might be a way for victims of rape or other forms of sexual abuse to ease their way back to becoming comfortable with sexual interactions with others. McArthur argues that if somebody has suffered a sexual trauma, it might be overwhelming for them to have sexual relations – or intimate relations more generally – with another human being. It might be more comfortable for them to experiment with intimate interactions with a robot first.

Actually, the case of Chuck seems to potentially fit this picture. As we noted above, there is a scene in the documentary *Hi, AI* in which Chuck is revealing to Harmony that he was the victim of sexual abuse as a child. Whether

[6] See the above-referred to debate between Richardson and Devlin.

or not Misselhorn is right that Chuck might be making himself into something thing-like in revealing sensitive details about his past to a robot, his case does seem to complicate the picture of somebody who wants to interact with a sex or companion robot as somebody who is eager to glorify or glamourize rape and sexual abuse.

Accordingly, we seem to need a more nuanced approach than one that likens all potential human users of sex robots or dolls to rapists and misogynists. A third interesting perspective to consider here is therefore a 'non-binary' answer to the question of whether humanoid sex and companion robots show disrespect for humans, which can be drawn from the work of Janna van Grunsven and Aimee van Wynsberghe (2019). The authors highlight two morally significant differences between the way humans are embodied and the way sex robots are, namely: unlike humans, sex robots have a lack of *bodily boundedness* and unlike humans, sex robots' emotional *expressivity and responsiveness is restricted.*

By "bodily boundedness" van Grunsven and van Wynsberghe specifically refer to the ways in which human beings are uniquely part of the world and separate from it, and the vulnerability to which this gives rise (2019: 8). To simplify their discussion, human bodies require interactions with their environment, for nutrition for example, but at the same time the "living bodily being" is an "autonomous being[]…who confront[s] the environment on their own terms" (ibid.). Human skin plays a central role in the experience of pleasure and pain as well as empathic experiences in general, and has a "social and existential significance" (ibid.). The extent to which robot skin shares or imitates these properties, the authors argue, has implications for the ways in which they will be perceived and the kind of ethical concerns we should have with respect to their use.

Van Grunsven and van Wynsberghe also argue that the emotional expressivity and responsiveness of robots is limited in ways in which humans are not. Human facial expressions, gestures, sounds and speech can express a vast array of emotions, attitudes, desires, etc., whereas what the sex robots of the present and near future will be endowed with are and will be far less expressive, in terms of both "scope and fineness of grain" (2019:: 11). The features that sex robots lack may be the very things that make them suited for certain therapeutic contexts, for example, for interventions involving those on the autism spectrum.

Return now to the quote from Davecat about how he does not want to treat the doll Muriel like a thing and how he won't do so. That was a claim about a doll, not about a robot. Suppose now, however, that there is a robot – perhaps a more advanced sex robot – that is able to simulate emotions and emotional behavior in an impressive way. This raises the third main question we will consider "no" and "yes" answers to.

3.3 Adequate Treatment of Humanoid Robots

Question: If a sex or love robot simulates emotional behavior, does this make a difference to how it is proper to treat it?

The topic of the moral status of robots in general, and humanoid robots in particular, as well as the consideration of how we should or should not treat them is increasingly becoming a contested subject (Gellers 2020; Birhane/Van Dijk 2021). It therefore serves us well to consider, once again, both "no" and "yes" answers to the question at hand. Let us begin with possible "no" answers, since these answers can be seen to be a bit more straightforward than any "yes" answers we may consider.

A first possible "no" answer is provided if we consider the above-mentioned purely instrumental view of technology. On this view, technology is seen as a value-neutral tool that we have designed and created for our use, and therefore robots are also mere technological devices designed and created for our use. Drawing upon Martin Heidegger's well-known (1977) "instrumental definition" of technology, David Gunkel (2018) explains that, from this point of view, the role and function of any kind of technology is to be utilized by human users as means to specific ends. This holds true for a simple hand tool such as a hammer, to a more sophisticated tool such as a Roomba vacuum cleaner to even more sophisticated technologies like a humanoid robot, e.g. Sophia the robot, the robotic copy of Hiroshi Ishiguro, or any advanced sex robot.

Much like we would utilize a Roomba vacuum cleaner to keep our floors clean, we may argue that a person may utilize a humanoid sex robot to attain the end of sexual gratification and companionship to combat loneliness. From this purely instrumental point of view, it would seem arbitrary to argue that technology requires proper treatment or any moral consideration (Bryson 2010; 2018; Müller 2021). Even if humanoid sex robots simulate emotional behavior (unlike a Roomba robot), this behavior is just that – *simulation*. It is not *real* emotion, and this simulated display of emotion is part and parcel of making it easier for people to attain certain ends with these robots because interacting with them will be more realistic. We can do whatever we want with tools we use to achieve whatever ends we want to achieve with those tools – or so it might be argued.

Another possible way to answer "no" to the question above is to consider one possible – and perhaps unexpected – interpretation of Mark Coeckelbergh's (2010) and Gunkel's (2018) "relational approach" to robot moral status. Briefly, the relational approach argues that the moral status of any robot will ultimately depend on the role it comes to play in our lives and on our relationships with the robot (Nyholm 2020: 195). This view could lead to the conclusion that we need not concern ourselves with how we treat these robots, if we also relate to the robots in a completely instrumental way. As such, if a robot does just play an instrumental role in our lives, then we do not have to be concerned with how we treat the robot in question, even if it might look and behave like a human being. (Müller 2021)

However, is not the whole point of creating humanoid robots that look human, and that simulate human emotions, to foster *human-like* relations with these robots, where therein lies also the potential for an *emotional* relation? The above-considered purely instrumental view of these robots apparently clashes with the very motivation behind their design and creation i.e., to be an imitation of a human partner. This point takes us to a possible "yes" answer to the question of whether humanlike appearance or behavior in a sex or love robot might make a difference to how we should behave around it, which applies Coeckelbergh's and Gunkel's relational approach in a more traditional way (Coeckelbergh 2010; Gunkel 2018; see also Loh 2019).

The very nature of these robots – that they look human and simulate human emotions – enables us to relate to them, and bond with them, in humanlike ways, as seen in the many examples we have considered. Therefore, if the robot comes to play a role in our life as that of an intimate companion, and we relate to it in this way, we should consider the robot to have some moral status or perhaps even similar moral status to that of a human to whom we can relate in the same way. It might seem morally problematic, for example, if Chuck started to perform violent actions towards Harmony after first interacting with the robot in the ways described above. The same could even be said about the example of Hanneke and Nadiah from the introduction above or about Davecat and Sidore. The ways in which they interact with these dolls seem to make it inappropriate for them to behave in apparently immoral ways towards them – at least there would be something symbolically problematic about it, i.e. it would symbolize something that is morally problematic in human-human relationships (Sparrow 2017; Nyholm 2020).

Another "yes" answer comes from John Danaher's (2020) "ethical behaviorism". Danaher argues that if the robot behaves the same way in which a human behaves (such as, in this case, by displaying human emotions), there is or might be a "performative equivalency" between the behavior of a robot and the behavior of a human. Danaher argues that we should therefore (i.e. because of this performative equivalency) consider the robot to have the same moral status as that of a human and treat the robot the same way in which we should treat a human being. Danaher thinks that in the human case, all we have to go on when it comes to deciding how to treat other human beings is how they behave and what is observable to us from the outside. If this is enough to help us determine how to treat other human beings from a practical point of view, it should also be enough, Danaher argues, to help us decide how to treat robots that behave like human beings.

A third possible perspective to consider in the context of "yes" answers to the question , relates to how our treatment of emotion-simulating robots might affect our own moral fiber (Friedman 2020). This is similar to Kant's (1996, 1997) well-known argument that in treating animals inhumanely, we become inhumane or cruel, and may become inhumane towards other human beings (Gerdes 2016; Darling 2016). In a similar vein, in the context of humanoid sex robots it might morally corrupt us if we treat them in apparently immoral ways. Many have written on this topic and approached it from differing perspectives.

Sparrow (2017: 549-471), for example, worries whether living out morally problematic rape fantasies by 'raping' robots would ultimately lead to a more aggressive behavior towards women, including sexual assault/rape incidents. Richardson, in her recent public presentations, makes similar claims.[7] Sparrow acknowledges that this is a contested topic, given "the claim that exposure to or enjoyment of representations of an activity makes people more likely to engage in that activity is heavily contested in the media effects literature" (Sparrow 2017: 470). Nevertheless, Sparrow thinks that 'raping' robots could possibly make some people more likely to rape women, thus leading to more women being raped:

> Sexual fantasy associates the imagining of behaviour with pleasure, which in turn associates the imagined behaviour with pleasure. Associating a fantasy of rape with sexual pleasure seems perilously close to a mechanism for

[7] See Richardson/Devlin 2021.

> Pavlovian conditioning for rape. At the very least, it might be expected to lower the barriers to rape by increasing the attractiveness of rape in the mind of the person who enjoys the fantasy. (Ibid.: 469)

Sparrow argues that for these reasons, it is possible to behave unethically in our interaction with sex or companion robots. Interestingly, however – as we noted above – Sparrow is of the opinion that while it is possible to behave inappropriately around such robots, he does not think that there could be such a thing as good behavior towards a robot (Sparrow 2020). In other words, if somebody followed Sparrow's advice and avoided any behaviors towards a robot that could be seen as symbolizing rape, violence, or any other form of immoral behavior, this unwillingness to treat robots in such ways could not be seen to be good or virtuous, according to Sparrow. It could only have the, so to speak, negative property of not being bad or vicious.

Kate Darling, in contrast, takes a different view. Darling (2016) argues that due to our social engagement with social robots, preventing the 'abuse' of these robots will protect our societal values and, thereby, be a form of good behavior. Darling specifically argues in the context of social robots that could be kept as pets and uses the example of a child kicking a robotic pet to get her point across:

> Given the lifelike behavior of the robot, a child could easily equate kicking it with kicking a living thing, such as a cat or another child. As it becomes increasingly difficult for small children to fully grasp the difference between live pets and lifelike robots, we may want to teach them to act equally considerately towards both. (Ibid.: 224)

Even in the context of adults interacting with these robots, she notes that "the difference between alive and lifelike may be muddled enough in our subconscious to warrant adopting the same attitudes toward robotic companions that we carry towards our pets" (ibid.: 224; for a fuller discussion, see Darling 2021). We could extend this argument to humanoid sex robots, given that due to their human-like form and display of human emotions, adults may form human-like bonds with them (Gordon/Nyholm 2022)[8].

[8] Of course, a question that often naturally arises here is whether the same could be said about violent pornography or video games. Does engaging with this type of media make people more likely to assault women, for example? It is difficult to say, since studies have not provided clear answers in this regard. Despite this, however, we do not think that sex robots and pornography/video games are necessarily comparable.

The issue of whether there is a proper way to treat these robots is, as this section indicates, quite delicate and nuanced. There are questions here not only about whether robots (e.g. humanoid sex robots) can *have* morally relevant abilities or properties. There are also questions about whether they can *imitate* such abilities or properties, as well as questions about whether they can *represent* or *symbolize* morally relevant abilities or properties. In short, can robots have, imitate, or represent/symbolize morally relevant properties, such as those having to do with the kinds of emotions human beings have in intimate or romantic relationship contexts? Even if we answer "no" to whether robots can have the morally relevant properties, this does not mean that anything goes in terms of how we should conduct ourselves around the robots. It might still be morally relevant that the robots can imitate or represent/symbolize morally relevant properties. That might be enough for it to be morally best to avoid certain ways of interacting with robots: it might even be enough for it to be good or right to interact with them in certain ways (Nyholm 2020: chapter 8).

4 Conclusion

Discussions about the topic of how human beings should interact with robots – including humanoid robots – have led to a broader debate surrounding the notion of rights for robots, given the connection between legality and morality (Gunkel 2018; Gellers 2020). Although what is legal is not always moral and what is moral is not always legal, there is a significant overlap (Asaro 2012). This robot rights-debate is just as complicated, and the need for nuanced discussion is just as strong (Schröder 2021). Although some scholars opt to see the consideration of robot rights in black and white, it is valuable to see the complexity and nuance that characterizes the topic.

Notably, the landscape of different possible views that can be taken about how human beings and robots should interact with each other – how human beings should behave around robots and how robots should be made to behave around people – is getting more and more complicated with every new publication coming out about this topic. In his book, *Robot Rights* (2018), for

This is because robots are embodied artificial agents, and not characters or avatars on a screen. Thus, we could argue that violent interactions with robots are more "realistic".

example, Gunkel mapped a lot of the existing literature that had been published up until that point. Based on an initial sketch of a visualization of this map by Danaher, Gunkel created an image where photographs of different researchers who have contributed to this debate are shown on a diagram that associates their pictures with different possible views that can be taken about this topic. Gunkel has continually been updating this picture, and he has been posting updated versions of the *robot rights map* on the social media site Twitter. Between 2018 when Gunkel's book came out and the present time – 2022 – Gunkel's robot rights map visualization is getting increasingly cluttered. And it is getting increasingly difficult to quickly characterize new contributions to this field in terms of where exactly they fit on the map.

We are of the opinion that attempts to map these discussions – such as the one above, which relates to emotional embodiment in humanoid sex and love robots in particular – are valuable for at least two reasons. One, they help to clarify what different views it is possible to take about the questions that arise when we are confronted with the reality of human beings interacting with robots that are made to look and behave like human beings. Second, these kinds of mappings help to illustrate the complexity of the ethical questions that arise in this context, and should, we suggest, lead us to try to adopt a nuanced understanding of the ethics of how human beings should interact with these kinds of robots. As Loh argues, we should not quickly dismiss different forms of human behavior around these robots as being evidence of "shortcomings", but should consider whether a commitment to valuing human diversity requires that we adopt a tolerant and open-minded approach to how people might want to interact with the robots of today and the robots of tomorrow.

Let us end by returning to where we started, with the scene from *Hi AI* in which Chuck is sitting down for a morning coffee together with Harmony. As we noted above, there is both something highly familiar about this scene – something cozy, if you will – and yet also something new and different, and potentially confusing and potentially even offensive to some. Can we really enjoy the simple pleasures associated with an intimate and romantic relationship with a robot, or would we be 'wasting' our emotions? Is this – and should this taken to be – offensive towards other human beings, who might feel that the suggestion is that they can be replaced by robots without minds or feelings? Or does the fact that robots can be made to look and behave in more humanlike ways than ever before put some pressure on us to avoid seemingly immoral behavior around these robots? We have considered some answers to these questions from the emerging and quickly growing literature on this topic. Our hope is that our discussion illustrates that these questions are not straightforward and that we need to approach them in a calm and careful way and not quickly dismiss the different perspectives that can be taken on this new and intriguing topic and part of human life.[9]

References

Asaro, P. M. (2012): "A Body to Kick, but Still No Soul to Damn: Legal Perspectives." In: P. Lin, K. Abney/G. A. Bekey (eds.), Robot Ethics: The Ethical and Social Implications of Robotics, Cambridge, MA: MIT Press, pp. 169-186.

Birhane, Abeba/van Dijk, Jelle (2021): "Robot Rights? Let's Talk about Human Welfare Instead." In: Conference on AI, Ethics, and Society. New York, February 2020.

Bryson, Joanna (2010): "Robots should be slaves." In: Y. Wilks (ed.), Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues, Amsterdam: John Benjamins, pp. 63-74.

Bryson, Joanna (2018): "Patiency is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." In: Ethics and Information Technology 10/1, pp. 15-26.

Carpenter, Julie (2016): Culture and Human-Robot Interaction in Militarized Spaces: A War Story. Emerging Technologies, Ethics and International Affairs, Burlington, VT: Ashgate.

Coeckelbergh, Mark (2010): "Robot Rights? Towards a Social-Relational Justification of Moral Consideration." In: Ethics and Information Technology 12/3, pp. 209-221.

Damiano, Luisa/Dumouchel, Paul (2018): "Anthropomorphism in Human-Robot Co-Evolution." In: Frontiers in Psychology 9, 468.

Danaher, J. (2019a): "The philosophical case for robot friendship." In: Journal of Posthuman Studies 3/1, pp. 5-24.

[9] Many thanks to Janina Loh and Wulf Loh, the participants of the "Social Robotics and the Good Life" workshop at Tübingen University, the "2021 Sex and Love with Robots" conference and the "Ethics, Anthropomorphism, and Robots" conference in Porto. Our work on this chapter is part of the research program Ethics of Socially Disruptive Technologies, which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

Danaher, John (2019b): “Building Better Sex Robots: Lessons from Feminist Pornography.” In: Yuefang Zhou/Martin H. Fischer (eds.), AI Love You, Berlin: Springer, pp. 133-148.

Danaher, John/McArthur, Neil (2017): Robot Sex: Social and Ethical Implications, Cambridge, MA: The MIT Press.

Darling, Kate (2016): “Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects.” In: R. Calo et al. (eds.), Robot law, Cheltenham: Edward Elgar, pp. 213-231.

Darling, Kate (2021): A Different Breed: What Our History with Animals Reveals about Our Future with Robots, New York: Henry Holt.

Davecat (2017): “THE DIG: Davecat, Married to a Doll”. *The Skin Deep*, October 12, 2021; https://www.youtube.com/watch?v=LiVgrHlXOwg.

Devlin, Kate (2018): Turned On: Science, Sex and Robots, London: Bloomsbury.

Devlin, Kate (2020): “The Ethics of the Artificial Lover”. In S. Matthew Liao (ed.), Ethics of Artificial Intelligence, New York: Oxford University Press, pp. 271-292.

Friedman, Cindy (2020): “Human–Robot Moral Relations: Human Interactants as Moral Patients of Their Own Agential Moral Actions Towards Robots”. In: Aurona Gerber (ed.), Artificial Intelligence Research, Berlin: Springer, pp. 3-20.

Gellers, Joshua (2020): Rights for Robots: Artificial Intelligence, Animal and Environmental Law, London: Routledge.

Gerdes, Anne (2016): “The Issue of Moral Consideration in Robot Ethics.” In: Acm Sigcas Computers and Society 45/3, pp. 274-279.

Gordon, John-Stewart/Nyholm, Sven (2021): “Ethics of Artificial Intelligence”. In: Internet Encyclopedia of Philosophy; https://iep.utm.edu/ethic-ai/.

Gordon, John-Stewart/Nyholm, Sven (2022): “Kantianism and the Problem of Child Sex Robots”. In: Journal of Applied Philosophy 39/1, pp. 132-147.

Gunkel, David (2018): Robot Rights, Cambridge, MA: MIT Press.

Gutiu, Sinziana (2016): “The Robotization of Consent”. In: Ryan Calo/Michael A. Froomkin/Ian Kerr (eds.), Robot Law, Cheltenham: Edward Elgar, pp. 186-212.

Heidegger, Martin (1977): The question concerning technology and other essays (W. Lovitt, Trans.), New York: Harper & Row.

Kant, Immanuel (1996): The Metaphysics of Morals, Cambridge University Press, Cambridge.

Kant, Immanuel (1997) Lectures on Ethics. Cambridge: Cambridge University Press.

Levy, David (2008): Love and Sex with Robots. London: Harper.

Loh, Janina (2019): Roboterethik: Eine Einführung, Frankfurt: Suhrkamp.

McArthur, Neil (2017): "The Case for Sex Robots". In: John Danaher/Neil McArthur (eds.), Robot Sex: Social and Ethical Implications, Cambridge, MA: The MIT Press, pp. 31-46.

Misselhorn, Catrin (2021): Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co, Berlin: Reclam.

Mullen, P. (2019): "Review: Hi, AI. Point of View Magazine." July 19, 2021; http ://povmagazine.com/articles/view/review-hi-ai.

Müller, Vincent C. (2020): "Ethics of Artificial Intelligence and Robotics". In: E. N. Zalta (ed.), Stanford Encyclopedia of Philosophy; https://plato.stanford.edu/entries/ethics-ai/.

Müller, Vincent C. (2021): "Is it Time for Robot Rights? Moral Status in Artificial Entities". In: Ethics and InformationTechnology 23, pp. 579-587.

Nyholm, Sven & Frank, Lily (2019): "It Loves Me, It Loves Me Not: Is it Morally Problematic to Design Sex Robots that Appear to Love Their Owners?" In: Techne 23/3, pp. 402-424.

Nyholm, Sven (2020): Humans and Robots: Ethics, Agency, and Anthropomorphism, London: Rowman & Littlefield International.

Nyholm, Sven (2022): "The Ethics of Humanoid Sex Robots". In: Brian D. Earp/Clare Chamber/Lori Watson (eds.), Handbook on the Philosophy of Sex and Sexuality, London: Routledge, pp. 574-585.

Nyholm, Sven/Danaher, John/Earp, Brian D. (2022): "The Technological Future of Love". In: Andre Grahle/Natasha McKeever/Joe Saunders (eds.), Philosophy of Love in the Past, Present, and Future, London: Routledge, pp. 224-239.

Nyholm, Sven/Frank, Lily Eva (2017): "From Sex Robots to Love Robots: Is mutual Love with a Robot Possible?" In: John Danaher/Neil McArthur (eds.), Robot Sex: Social and Ethical Implications, Cambridge, MA: The MIT Press, pp. 219-244.

Richardson, Kathleen (2015): "The Asymmetrical 'Relationship': Parallels Between Prostitution and the Development of Sex Robots". In: SIGCAS Computers & Society 45/3, pp. 290-293.

Richardson, Kathleen (2016): "Technological Animism: The Uncanny Personhood of Humanoid Machines". In: Social Analysis 60/1, pp. 110-128.

Richardson, Kathleen/Devlin, Kate (2021): "Sex, Robots, and Artificial Intimacy", a 2021 debate featuring Kathleen Richardson and Kate Devlin, moderated by Lily Frank, October 13, 2021; https://www.youtube.com/watch?v=kDC82yhpZ7A.

Riek, Laurel D./Rabinowitch, Tal-Chen/Chakrabarti, Bhismadev/Robinson, Peter (2009): 4th ACM/IEEE International Conference on Human Robot Interaction, California, March 9-13.

Royakkers, Lamber/Van Est, Rinie (2015): Just Ordinary Robots: Automation from Love to War, Boca Raton, FL: CRC Press.

Ryland, Helen (2021): "It's Friendship, Jim, but Not as We Know It: A Degrees-of-Friendship View of Human-Robot Friendship." In: Minds and Machines 31/3, pp. 377-393.

Scheutz, Matthias (2012): "The Inherent Dangers of Unidirectional Emotional Bonds." In: P. Lin/K. Abney/G. A. Bekey (eds.), Robot Ethics: The Ethical and Social Implications of Robotics, Cambridge, MA: The MIT Press, pp. 205-221.

Schoorl, John (2020): "Hanneke Wetzer zocht een lockdownbuddy en vond Nadiah op sexpopwebshop.nl". In: de Volkskrant, October 12, 2021; https://www.volkskrant.nl/mensen/hanneke-wetzer-zocht-een-lockdownbuddy-en-vond-nadiah-op-sexpopwebshop-nl~b650e847/.

Schröder, Wolfgang M. (2021): "Robots and Rights: Reviewing Recent Positions in Legal Philosophy and Ethics". In: J. von Braun/M. S. Archer/G.M. Reichberg/M. Sánchez Sorondo(eds.), Robotics, AI, and Humanity, Cham: Springer, pp. 191-203.

Sparrow, Robert (2017): "Robots, Rape, and Representation". In: International Journal of Social Robotics 9/4, pp. 465-477.

Sparrow, Robert (2020): "Virtue and Vice in Our Relationships with Robots: Is There an Asymmetry and How Might it be Explained?" In: International Journal of Social Robotics 13, pp. 23-29.

Turkle, Sherry (2007): "Authenticity in the Age of Digital Companions". *In: Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 8/3, pp. 501-517.

Turkle, Sherry (2012): Alone together: Why we expect more from technology and less from each other, New York: Basic Books.

Turner, Jacob (2019): Robot Rules: Regulating Artificial Intelligence, Cham: Palgrave Macmillan.

van Grunsven, Janna/van Wynsberghe, Aimee (2019): "A Semblance of Aliveness: How the Peculiar Embodiment of Sex Robots Will Matter". In: Techne 23/3, pp. 290-317.

Wendland, Karsten (2020): "Dass Roboter uns Emotionen vorgaukeln kann sehr wichtig sein. Im Gespräch mit Janina Loh". In: Selbstbewusste KI; h ttps://doi.org/10.5445/IR/1000125862.

Wetzer, Hanneke (2020), "Living Alone (Together)". In: Hannekewetzer.com, September 22, 2021; https://www.hannekewetzer.com/artportfolio/living alonetogether.

Contributors

Jacqueline Bellon: Jacqueline Bellon is a research associate at University of Tübingen (IZEW). Research interests include philosophy of technology, epistemology, history and philosophy of science and psychology, history of ideas, social theory, intersections of philosophy, sciences, and arts.

Leonie N. Bossert: Leonie N. Bossert is Post-Doc at the International Centre for Ethics in the Sciences and Humanities (IZEW) and at the Chair Ethics, Philosophy and History of the Life Sciences at the University of Tübingen. She studied Landscape Ecology and Nature Conservation at the Universities of Greifswald and Copenhagen and received her PhD at the University of Tübingen (2021) for a dissertation on an animal ethical theory of Sustainable Development. She is working as a researcher and lecturer with a focus on Sustainability Development, Animal, Environmental and Conservation Ethics, Philosophy of Nature and Human-Animal Studies.

Cordula Brand: Cordula Brand is managing director and scientific coordinator of the International Centre for Ethics in the Sciences and Humanities (IZEW) at the University of Tübingen. She studied philosophy at University Duisburg-Essen and received her PhD at the University of Tübingen, followed by a PostDoc position at the DFG research training group "Bioethics" and a Humboldt-Fellowship at Keio University, Tokyo. In research and teaching she deals with metaethical questions concerning moral psychology, theories of intersubjectivity and personhood as well as concepts of moral reasoning. Furthermore, she is interested in research and organizational ethics.

Charles Ess: Charles Ess is Professor Emeritus, Media Studies, University of Oslo. He works across the intersections of philosophy, computing, applied ethics, comparative philosophy and religious studies, and media studies, with emphases on internet research ethics, Digital Religion, virtue ethics, social robots and AI. Ess has published extensively on ethical pluralism, culturally-variable ethical norms and communicative preferences in cross-cultural approaches to Information and Computing Ethics, and their applications to everyday digital media technologies.

Lily Eva Frank: Lily Eva Frank is an Assistant Professor of Philosophy and Ethics in the Department of Innovation Science at the Eindhoven University of Technology. Her research focuses on the ethics of socially disruptive technologies, including robotics, emerging biotechnology, and technologies of love and sexuality.

Cindy Friedman: Cindy Friedman is a PhD Candidate at Utrecht University. Her research is a part of the Ethics of Socially Disruptive Technologies research programme, and is focused on the ethics of humanoid robots. She has a broad interest in the ethics of social robots and human-robot interaction. She completed her Masters at the University of Pretoria in South Africa, where she conducted research on the ethics of sexbots.

David Gunkel: David J. Gunkel is an award-winning educator, scholar, and author, specializing in the philosophy and ethics of emerging technology. He is the author of over 90 scholarly articles and book chapters and has published fifteen internationally recognized books, including *Thinking Otherwise: Philosophy, Communication, Technology* (Purdue University Press 2007), *The Machine Question: Critical Perspectives on AI, Robots, and Ethics* (MIT Press 2012), *Of Remixology: Ethics and Aesthetics After Remix* (MIT Press 2016), and *Robot Rights* (MIT Press 2018). He currently holds the position of Presidential Research, Scholarship and Artistry Professor in the Department of Communication at Northern Illinois University (USA).

Karen Lancaster: Karen Lancaster is a 4th year PhD Philosophy student at the University of Nottingham. Her thesis explores ethical issues in the implementation of care robots in residential homes for elderly people, with specific focus on the nature of care, dignity, and consent. Karen has previously published articles on both sex robots and care robots, and she has undertaken a variety of outreach work, teaching various philosophical topics in primary schools, secondary schools, colleges, and at university level. Her philosophical interests include social, political, ethical, and legal philosophy, particularly issues arising from our use of emerging technologies.

Janina Loh: Dr. Janina Loh (née Sombetzki) is an ethicist (Stabsstelle Ethik) at Stiftung Liebenau in Meckenbeuren on Lake Constance. They published the first German *Introduction to Trans- and Posthumanism* (Junius 2018) and an *Introduction to Robot Ethics* in German language (Suhrkamp 2019). They habilitate on a *Critical-Posthumanist, Queer, Relational (Inclusive) Ethics of Companionship.* Loh's main research interests lie in the field of trans- and posthumanism (especially critical posthumanism), robot ethics, feminist philosophy of technology, responsibility research, Hannah Arendt, theories of judgement, polyamory, and ethics in the sciences.

Wulf Loh: Dr. Wulf Loh is Assistant Professor at the Int. Center for Ethics in the Sciences and Humanities (IZEW) at the University of Tuebingen, Germany. At the Center, he directs the research focus "AI and Robotics" and is PI of various technology development projects. His areas of expertise are within Applied Ethics (AI Ethics, Media and Information Ethics, Privacy, Robot Ethics, Human Machine Interaction), Social Philosophy (Critical Theory, Practice Theory, Social Ontology), Political Philosophy (Digital Public Spheres, Digital Civil Disobedience, International Political Theory), and Philosophy of Law (Constitutional Theory, Legitimacy and Legality, Philosophy of International Law).

Sven Nyholm: Sven Nyholm is an associate professor of philosophy at Utrecht University in the Netherlands. His publications include the books *Revisiting Kant's Universal Law and Humanity Formulas* (De Gruyter 2015), *Humans and Robots: Ethics, Agency, and Anthropomorphism* (Rowman & Littlefield International 2020), and *This is Technology Ethics: An Introduction* (Wiley-Blackwell 2023).

Tom Poljanšek: Tom Poljanšek is post-doc and assistant to the chair of theoretical philosophy at University of Göttingen. Research interests include theories of perception, phenomenology, social ontology, philosophy of technology, and aesthetics.

Thomas Potthast: Thomas Potthast holds the Chair for Ethics, Philosophy and History of the Life Sciences, and is Director of the International Centre for

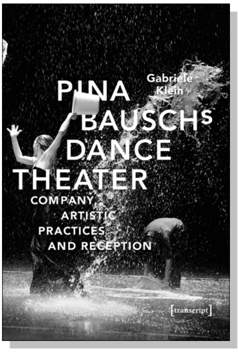
Ethics in the Sciences and Humanities (IZEW) at the University of Tübingen. He studied biology and philosophy at Freiburg, Germany, and received his PhD at Tübingen (1998), followed by a PostDoc at the Max-Planck-Institute for the History of Science, Berlin, (1998-2001) and a Humboldt-Fellowship at the University of Madison-Wisconsin/USA (2002). His research and teaching addresses ethical and epistemological dimensions of science, including conceptual and practical questions of interdisciplinary and transdisciplinary research at the science-society-interface, with a special focus on human-nature relations, (bio)technologies, and sustainable development.

Anna Strasser: After postdoctoral positions as scientific researcher and coordinator in Freiburg (Center for Cognitive Science) & Berlin (Berlin School of Mind and Brain) and a visiting fellowship at the Center for Cognitive Studies at Tufts University with Daniel Dennett, I founded the DenkWerkstatt Berlin and work now as an independent, freelance philosopher in Berlin. Since autumn 2020, I am an associate researcher of the Cognition, Values, Behaviour (CVBE) research group at LMU-Munich (philosophy). In my research, I focus on topics concerning social cognition at the intersection of philosophy, psychology, and AI. Currently, I am primarily interested in how a philosophical framework can capture the varieties of phenomena in social cognition. To this end, I question whether standard notions are too restrictive and examine to what extent so-called minimal approaches can contribute to a solution. In addition, I investigate to what extent artificial systems may qualify as a new type of a social agent.

Imke von Maur: Imke von Maur is philosopher, working as a postdoctoral researcher at the Institute of Cognitive Science of Osnabrück University. In her PhD thesis she developed an approach of the epistemic relevance of emotions from a decidedly socio-critical stance. Her current work focuses on what it means and requires to understand complex phenomena, such as the climate crisis, and the implications for (higher) education, and on the ethics of and narratives around the role of AI and technology in and for society.

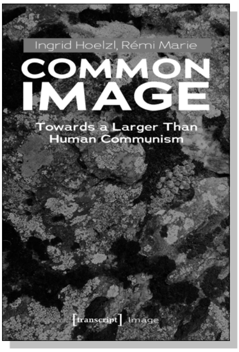
Eva Weber-Guskar: Eva Weber-Guskar is a professor of ethics and philosophy of emotions at the Ruhr University Bochum. Currently, she is working on ethical questions in dealing with AI, especially Emotional AI, and on temporal aspects in theories of the good life. She is a PI in the interdisciplinary project “INTERACT! New forms of social interaction with intelligent systems” at the RUB. Prior to this, she held positions in Berlin, Vienna, Zurich and Erlangen and spent one year as a visiting scholar at New York University. Her second book is on human dignity (*Würde als Haltung*, Mentis 2016) and her first book (PhD) is on understanding emotions (Die Klarheit der Gefühle. Was es heißt, Emotionen zu verstehen, De Gruyter 2009). Furthermore, she is a co-founder of PhilPublica, an initiative to foster academic philosophy in the public domain.

Cultural Studies


Gabriele Klein
PINA BAUSCHS DANCE THEATER
COMPANY, ARTISTIC PRACTICES AND RECEPTION
[transcript]

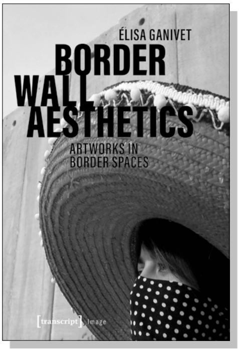
Gabriele Klein
Pina Bausch's Dance Theater
Company, Artistic Practices and Reception

2020, 440 p., pb., col. ill.
29,99 € (DE), 978-3-8376-5055-6
E-Book:
PDF: 29,99 € (DE), ISBN 978-3-8394-5055-0


Ingrid Hoelzl, Rémi Marie
COMMON IMAGE
Towards a Larger Than Human Communism
[transcript] image

Ingrid Hoelzl, Rémi Marie
Common Image
Towards a Larger Than Human Communism

2021, 156 p., pb., ill.
29,50 € (DE), 978-3-8376-5939-9
E-Book:
PDF: 26,99 € (DE), ISBN 978-3-8394-5939-3



Elisa Ganivet
Border Wall Aesthetics
Artworks in Border Spaces

2019, 250 p., hardcover, ill.
79,99 € (DE), 978-3-8376-4777-8
E-Book:
PDF: 79,99 € (DE), ISBN 978-3-8394-4777-2

All print, e-book and open access versions of the titles in our list are available in our online shop www.transcript-publishing.com

Cultural Studies