Social Robotics and Synthetic Ethics: A Methodological Proposal for Research

Bako Rajaonah, Enrico Zio

The final publication is available at http://link.springer.com: Rajaonah, B., Zio, E. Social Robotics and Synthetic Ethics: A Methodological Proposal for Research. Int J of Soc Robotics (2022). https://doi.org/10.1007/s12369-022-00874-1

Title Page

Title

Social Robotics and Synthetic Ethics: A Methodological Proposal for Research

Authors

Bako Rajaonah[a], Enrico Zio[bc]

Affiliations

[a]Laboratory of Industrial and Human Automation control, Mechanical engineering and Computer Science (LAMIH UMR CNRS, 8201), Université Polytechnique Hauts-de-France, Valenciennes, France

[b]Centre de Recherche sur les Risques et les Crises (CRC), MINES ParisTech/PSL Université Paris, Sophia Antipolis, France.

[c]Department of Energy, Politecnico di Milano, Milan, Italy.

E-mail addresses & ORCID

Dr Bako Rajaonah: bako.rajaonah@uphf.fr; ORCID: 0000-0001-8070-9308
Professor Enrico Zio: enrico.zio@mines-paristech.fr; enrico.zio@polimi.it; ORCID: 0000-0002-7108-637X

Corresponding author

Dr Bako Rajaonah
bako.rajaonah@uphf.fr
LAMIH, Université Polytechnique Hauts-de-France
Campus Mont Houy, Valenciennes, F-59313
Phone number: +33 27 51 14 91

Abstract

This paper outlines a methodology to analyze the coconstruction of ethical interactions between humans and social robots. Indeed, in view of the aging populations of many societies worldwide, such robots could be useful for helping caregivers in many ways. Nevertheless, ethical robots are a precondition for this possibility. Based on the concept of ethical know-how posited by Francisco Varela, the discipline of synthetic ethics, and an existing model of intrinsically moral robots, this methodology employs an experimental user-centered research design for a qualitative study of interactions between elderly people and social robots. The goal of the methodology is to provide information concerning how interactions between elderly people and social robots could produce ethical expertise on both sides. The methodology relies on the techniques of nonparticipant observation, focus group discussion, and questionnaires. This proposal is expected to be of interest to individuals who are involved in the design of ethical social robots for the care of vulnerable people.

Keywords

Ethical know-how; Synthetic ethics; Social robot; Elderly care; Qualitative research

The final publication is available at http://link.springer.com: Rajaonah, B., Zio, E. Social Robotics and Synthetic Ethics: A Methodological Proposal for Research. Int J of Soc Robotics (2022). https://doi.org/10.1007/s12369-022-00874-1

Social Robotics and Synthetic Ethics: A Methodological Proposal for Research

Ikso moved cautiously toward the old lady, sat down at her side, and put her hand on hers. A pale smile lit the lady's face. Ikso had perceived her anxiety and wished to ease it; she was all white and metal, and the silence was disturbed only by the faint sound of machines purring when a little cry was heard from nearby. Without getting up, Ikso checked her screen and noticed that the patients in adjoining rooms were sleeping quietly. She nevertheless asked the night supervisor to inspect each room on the floor. Igréko burst into the room a few minutes later and said that their colleague Zédo had disappeared and was unreachable. Zédo was in charge of the care and supervision of the occupant of a nearby room. What was not seen right away was that the wire connecting the person to life support had been unplugged. The cry heard before was that of a death rattle. They understood that Zédo had euthanized the old man. His grandson was somewhat astonished when he received the certificate of death. Indeed, he did not remember this grandfather who had entered the nursing home three years before. For Zédo, he was found on the upper floor disconnecting other wires. Ikso and Igréko no longer trusted Zédo to take care of humans. They thus decided that he would do nothing else but cooking for the residents […] ([1], p. 53).

1 Motivation

This paper proposes a methodology for examining the coconstruction of ethical interaction between humans and social robots from the perspective of synthetic ethics. Robots have been a popular subject since the early days of science fiction, and social robots are no exception to this rule. Hence, it is not surprising that science fiction has been used to examine to deployment of social robots and now addresses related scientific issues such as ethics [2, 3]. The short story discussed above is thus a relevant preamble to work tackling the ethical issues related to using robots for the care of elderly people.

A social robot is an autonomous or semiautonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with whom the robots are intended to interact [4, p. 592]. Shaw-Garlock employed the term "affective humanoid social robots" for "robots that are designed to interact with humans on an emotional level" ([5], p. 250), which Ikso, Igréko and Zédo could be: they interact with humans on the **human emotional level**. Equipped with artificial emotional empathy, such robots are able to recognize facial, gesture, and/or

vocal emotional information expressed by their human partners, to select and deploy an appropriate response so that humans can understand the response in terms of the robot expressing empathy for their feelings [6].

The goal of robots such as Ikso, Igréko and Zédo could be to replace people who take care of elderly individuals, at certain times and under certain circumstances, but without taking the place of human caregivers, instead assuming a role to which Dumouchel and Damiano referred as that of a substitute. According to those authors, a substitute is a social actor characterized by at least the following interrelated features: a capacity for social autonomy, a capacity for performing more than one activity, and a capacity for exercising authority over the human partner; those authors made it clear that "it is neither required nor expected of the substitute that he will take the place of the person on whose behalf he is acting" ([6], p. 32). These capacities enable the substitute to have some room to maneuver and initiate or rearrange activities accordingly during the course of interaction with the full approval of the human partner. Let us take the example of Robot & Frank, a science-fiction movie from 2012 written by Christopher Ford and directed by Jake Schreier. Frank is an elderly individual, and Robot is a "health care aid programed to monitor and improve his physical and mental health."[1] When Frank says that he does not like gardening, Robot firmly suggests another physical activity, namely, walking, a more moderate physical activity than gardening. And so they walked. This kind of authority exerted by Robot might be technically possible in the future, but the very idea may currently seem impossible to accept. However, what about subsequent generations? What about those who are already very sensitive to social influence through social networks? It is likely that the very ideas of obedience and authority may change over time and may depend on the generation in question. However, the concrete consequence of such a situation is that such robotic substitutes could be empowered to be true social partners and could therefore become useful and valuable for vulnerable people such as lonely elderly people.

Despite these benefits, such substitutes are not without challenging issues, such as the morality-related questions raised by the story of Ikso, Igréko and Zédo, for instance, the normalization of the abandonment of elderly people or issues regarding responsibility for decision-making with respect to euthanasia. The present work does not deal with those societal issues but rather focuses

[1] Script of the movie Robot & Frank: https://www.scripts.com/script/robot_and_frank_17060 (Accessed November 26, 2021)

on the morality of substitutes. Morality regulates our social behavior in terms of what is right and wrong [7]; if such robots are to be our social partners in the future, then the question of their morality is as important as it is in relationships among humans.

However, the viewpoint of neuroscientist and philosopher Francisco Varela deserves to be considered: inspired by Eastern teaching traditions, he suggested that ethical behavior could also be seen from a nonmoralistic perspective; more precisely, "why should one conflate ethical behavior with judgment?" ([8], p. 4)[2]. According to Varela, most philosophers and scientists have focused on know-what and neglected the skilled behavior that is called ethical know-how. He did not deny the importance of deliberation and analysis before behaving, but he did relegate these activities to situations in which one has no prior ethical expertise: most of the time, one simply performs ethical actions, that is, actions that he or she knows are correct for the purposes of immediate coping. Referring to Mencius ("an early Confucian from around the fourth century B.C.E.," p. 26), Varela noted that: "people actualize virtue when they learn to extend knowledge and feelings from situations in which a particular action is considered correct to analogous situations in which the correct action is unclear" (p. 27). This actualization is what Varela called ethical know-how. Know-how in general is learned from our personal history of interactions with others, emerging from the "recurrent sensorimotor patterns that enable action to be perceptually guided" (p. 12) in the "perceiver-dependent world" (p. 13). "It is also clear that we can add responding to the needs of others to our list of skilled behaviors" (p. 23). Furthermore, "*et que les conclusions précédentes valent également pour l'étude des actes et du savoir-faire éthiques*" ([9], p. 44)[3]. The term "conclusions" refers to Varela's claim that "my main point is that most of our mental and active life is of the immediate coping variety" ([8], p. 19). In other words, ethical know-how is another form of know-how that humans apply spontaneously in everyday life.

Varela's view is relevant and innovative with respect to the present work concerning social robots and synthetic ethics: both ethical know-how and synthetic ethics are grounded on the concept of embodied cognition, which allows researchers in the field of robotics to address the ethicality of a social robot due to its capacity for acquiring know-how through interactions with a specific

[2] In the French version of the book, Varela asked: "*Pourquoi confondre le comportement éthique et le jugement moral*?" ([9], p. 18)

[3] The French version slightly different from the English version, which is why the present authors preserve the citation in French. In English, the entire sentence is "It is also clear that we can add responding to the needs of others to our list of skilled behaviors without doing violence to our concept of ordinary life"

environment. Researchers do not have to puzzle over which moral theory is the more appropriate in the context of a social robot. Indeed, as pointed out by Wallach, this question has not yet been solved with respect to human morality despite the "three-thousand-year inquiry of moral philosophers into whether any one ethical theory is adequate for capturing the breadth and complexity of human moral considerations […]" ([10], p. 467).

The present work aims to propose the outline of a methodology to study the feasibility of social robots developing ethical know-how without their artificial intelligence having to compute a set of moral rules when they interact with humans during their everyday activities. Indeed, according to Dumouchel and Damiano, the substitute's fundamental characteristic of social autonomy that enables the mutual coordination of emotions, intentions, and behaviors with human partners is not compatible with the principle of preset moral rules: the latter would constrain the substitute's ability to maneuver and hinder the emergence of creative interactions, and yet the emergence of creative interactions during the coevolution of humans and social robots in the near future is an opportunity for ethical innovation, which can be examined experimentally via the new discipline of synthetic ethics [6].

The following section defines the concepts that underlie both synthetic ethics and the proposed methodology for an exploratory study of the development of ethical behavior between humans and social robots. It also briefly presents an existing model of intrinsically moral agents on which the proposed methodology relies. Section 3 outlines this methodology, and the final section concludes in light of the risks that society could encounter with the introduction of social robots for the care of elderly people.

2 Synthetic Ethics and Underlying Concepts

All concepts that underlie synthetic ethics and are important in the design of the methodology proposal are highlighted in bold type in this section.

The **synthetic approach** aims to apply artificial systems to understand the hidden mechanisms that are inaccessible to our perception and understanding and that underlie the complex behavior that emerges from interaction between living organisms and their environment [11, 12]. The approach consists of first building a model of the behavior using an artificial agent and then analyzing the observed behavior rather than vice versa (i.e., not of analyzing a behavior to build an explanatory model of it). This sequence is why Dumouchel and Damiano believe that the synthetic approach is

appropriate for an experimental study of the development of ethical behavior between substitutes and their human partners, and, according to those authors, the approach can also provide knowledge regarding the growth of morality in humans [6].

Seibt proposes the approach of **integrative social robotics** to focus on the interaction between humans and robots rather than on the robot alone to determine what the robot can and should do during such interactions [13]. Damiano agrees with Seibt when she considers both the **coevolution** of humans and robots in the near future and the **coconstruction** of their ethics with respect to the new forms of interaction that emerge through their **self-organization** [14]. Indeed, self-organization is characterized by the spontaneous **emergence** of high order properties among system components, which cannot be reduced to the properties of each component [15].

The phenomenon of emergence is also important in the approach of the **joint cognitive system**. Like Seibt's conception of the human-robot interaction unit, Hollnagel and Woods (2005) consider the joint system of human and machine to be the unit of analysis [16], and like Dumouchel and Damiano's views concerning the coevolution of humans and robots, Hollnagel and Woods emphasize the importance of considering the **coagency** of humans and machines to achieve a **common goal**: the overall performance emerging from the joint system. In the present work, the common goal of both robots and their users is to acquire ethical know-how regarding their interactions.

In all these approaches, the role of **the environment** is central because humans and artificial agents interact with the environment, and the environment contributes to shaping those parties individually, along with their interactions and coagency.

Embodied cognition means that cognitive processes and bodily interactions with the environment are deeply interdependent, the starting assumption being that the mind exists through bodily emotional, perceptive, and motor experiences [17, 18]. Embodiment refers concretely to the coupling of both low- and high-level cognitive processes with neural sensory-motor structures (e.g., [19]) through the transformation of perceptual information into a motor format (e.g., [20]) and through the formation of brain motor patterns that are activated when movements are replayed, observed, or imagined (e.g., [21]).

It is precisely this potential of the embodied cognition approach to deal with complex and high-level cognitive processes by learning, without having to specify all the rules from the beginning,

that interests researchers in social robotics [22]. For instance, Breazeal et al. [23] applied the embodiment approach to equip robots with **social cognition skills**; their robot develops the capability of mind reading through the observation and simulation of behavior.

In the same context, based on the bottom-up approach of moral machines, Balkenius et al. [24] proposed endowing robots with the capability to develop their own morality through sensory-motor interactions with other agents and the environment. Their model robot is able to **learn to distinguish between right and wrong behavior** through observations of the emotions expressed by others via the body and face and through interaction with those individuals and the environment. Specifically, the robot is capable of observational learning, understanding the goals of others, and recognizing causal relationships between actions and their consequences. Balkenius et al. grounded their work on the assumption that moral behavior is connected to social emotions. The social emotions with which these authors experimented were oriented on an orthogonal framework with hope/pride and fear/shame/embarrassment on one axis and frustration and relief on the other. To operationalize these emotions with regard to morality, they are considered to be a form of preparation for the expected reactions of others: hope vs. shame arises when expectations of positive vs. negative events are fulfilled, while relief vs. frustration arises when expectations of negative vs. positive are not fulfilled.

Like Varela, Balkenius et al. grounded their work in the embodied cognition approach, but the difference is that the latter authors refer explicitly to morality, whereas Varela situated his philosophical reflection in a nonmoralistic framework. Indeed, as mentioned previously, most of the time in everyday life, one behaves ethically without prior moral judgment, which Varela called spontaneous coping based on **“ethical expertise”** or **“ethical know-how,”** as opposed to intentional, rational judgment related to “ethical know-what” in unfamiliar situations [8, 9].

Concretely, in the model of Balkenius et al., actions are perceived to be moral or not moral through the emotions expressed by the human partner. What is called moral behavior by Balkenius et al. is called ethical behavior by Varela. This concept represents the learning capacity of their model and is relevant for the present work because that model suits the goal of the proposed methodology, i.e., the task of examining the colearning of ethical interactions by humans and social robots in accordance with the discipline of synthetic ethics.

The focus of the present work is human-robot interaction, but it takes the **sociotechnical environment** into account: **stakeholders** involved in the design and development of, as well as coexisting with, social robots in the near future should also be included in the study because the role of societal factors cannot be neglected when understanding the social acceptability of such robots (see, for example, the work of Winkle et al., who conducted a survey to examine the social acceptance of socially assistive robots [25]).

Finally, **well-being** is omnipresent in most ethics guidelines for AI-based systems (e.g., [26–28]): such systems "must permit the growth of the well-being of all sentient beings" ([26], p. 8). Therefore, the well-being of the humans who are intended to interact with social robots should be investigated systematically. Subjective well-being is formally defined by the "Organisation for Economic Co-operation and Development" (OECD) as good mental states, including the evaluations that people make of their lives, as well as their affective reactions to their experiences, and this organization has published guidelines for measuring subjective well-being [29].

3 Methodological Proposal

The goal of this proposal is to provide an approach to experimentally examine the ways in which ethical know-how emerges and develops from interactions between humans and social robots. The proposal takes the form of a user-centered research design for a sociotechnical system based on nonparticipant observation and focus group discussions with targeted users and stakeholders. The context is the care of elderly people by robotic substitutes in a nursing home. In accordance with the synthetic approach, the experiment is based on the use of an artificial agent, "a scientific instrument" ([6], p. 55); here, this agent is the model described by Balkenius et al. [24].

It is absolutely necessary to emphasize the fact that this proposal is only an outline of a methodology that could be implemented in a real project. Such a project is necessarily transdisciplinary and encompasses many kinds of stakeholders involved in the design, development, deployment, and use of the social robot under study: as such, once academic partners and stakeholders agree with the goal of study, each of them must have a say in the methodology (such as in [30, 31]).

This section describes the procedure for gathering qualitative data to obtain information concerning perceived acceptable interactions between humans and robots.

3.1 Hypothesis

First, let us note that we use the term "acceptable" in this methodological context instead of "ethical" or "moral" to avoid biasing the results. The underlying assumption is that interactions that are consensually considered to be being socially acceptable at the level of the participants might reasonably be conceived of as ethical, that is, as not forbidden: "On the one hand, social norms determine what kind of behavior of amoral agent is acceptable […]. On the other hand, social norms also attribute rights and values to objects of action, indicating what kind of behavior towards such objects is morally appropriate" ([32], p. 296). The word acceptable here could also be related to the term "morally praiseworthy" expressed by Wallach when explaining bottom-up approaches to the development of agents' morality ([10], p. 467).

The main hypothesis of this proposal is that many interactions between human and robotic agents favor the emergence of behaviors and interactions that are progressively mutually accepted, the underlying assumption being that if the behaviors of one agent are accepted by the other, then these behaviors are considered correct by this other. The acceptance or nonacceptance of a behavior is inferred by the observers (participants in focus group discussions) through the social emotions expressed by the agent depending on the fulfillment of their expectations, that is, hope, shame, frustration, and relief in accordance with the model developed by Balkenius et al.

3.2 Research Questions

To obtain valuable information from the experimental study, the research questions of such a study could be as follows:

—Which interactions are accepted and which are not accepted by humans or robots? Robots are considered because a social robot who has acquired certain ethical know-how could "refuse" to perform a particular behavior.

—In which contexts are those interactions accepted or not accepted?

— Can interactions that are deemed acceptable by the interactors be considered to be ethical interactions?

—Can interactions that promote well-being for humans be considered to be acceptable?

—Are interactions unethical when they are not accepted by either humans or robots?

— Are interactions unethical when they do not promote well-being for the human agent during and after interactions with a robot?

—Can invariance be observed in these interactions, such that it could be said that there is invariance in ethical interaction?

—If this invariance can be observed, how long does it take to emerge?

—Can moral behaviors (in the sense in which people usually understand morality, i.e., in terms of right and wrong, good and evil), acceptable interactions, and well-being be conceptually related in the particular context of elderly people interacting with social robots?

—For further generalizations, are invariant ethical features of know-how experimentally observable? If so, does this invariance converge with existing moral principles?

— Regarding the model developed by Balkenius et al., what other emotions could be implemented, for example, with regard to well-being (see § 3.5)? Would it be possible in the future to make the model capable of detecting psychological states such as relaxation, unease, awkwardness, tiredness, anxiety, etc.?

3.3 Site

The research should take place in a nursing home or, at least, a nursing-home-like place, that is, a location that matches the conditions and facilities of a nursing home, because it will be easier and less pervasive to equip a nursing home with social robots and recording equipment than to equip a home with such material and because a large proportion of participants (the nursing staff) will already be on site.

3.4 Participants

—The social robot is considered to be a participant (PR). At least two assistive, social robots based on the model developed by Balkenius et al. are necessary to enable the following units of interaction: human-robot, humans-robot, human-robots, humans-robots, and even robot-robot. This variety is important because the configuration should not be limited to the classic unit of human and machine but should rather be closer to natural social networks and should provide an opportunity to observe more novel forms of interaction.

—The second kind of participant (PH) is the targeted user of social robots such as Ikso, Igréko, and Zédo, that is, the elderly person. However, this definition is not yet relevant: future targeted users are not yet elderly. To allow for generalization of the results, people of a generation younger than 65 years old should be combined with people 65 years and older, for example, people approximately 40 years old and their parents, or even the children of those who are approximately forty. PHs should also include nurses who work on the site and interact with both PRs and other PHs. Various kinds of PH can provide more variety with respect to the forms of interaction. Including children might shed light on the contents of both obedience and authority in the context of interaction with social robots, which may change over time and depend on the generation involved (see § 1), which could be one expression of ethical innovation, among others. Moreover, examination of the intention to benefit from such robots in the future can be more relevant.

—The last kind of participant pertains to the stakeholders involved in the design of social robots for a nursing home: engineers and researchers, gerontologists, nurses, jurists, representatives of insurance companies, providers of health and social services to elderly people, etc.

All stakeholders, including PHs, should participate in focus group discussions with the main goal of analyzing and interpreting the observed interactions.

3.5 Techniques for Data Collection

The techniques used to collect data are nonparticipation observation, focus group discussion, and questionnaires.

- **Nonparticipant observation** is a technique used by researchers to examine the subjects of their study without taking part in the examined situation; to avoid participants modifying or improving their behavior because they know that they are being observed, "researchers normally observe a number of similar situations, over a period of time" ([33], p. 518). **Focus group discussion** is among the techniques employed during user-centered systems design processes (e.g., [34]). The principle of focus group discussion is to assemble people who are relevant to a specific subject and to encourage them to interact with each other, exchange viewpoints and comment on each other's experience [35, 36], for example, a group of future end users could be consulted to assess a new technology or service. Focus group discussions allow the researcher "to obtain data regarding ideas, attitudes, understanding and perceptions" ([37], p. 69); such discussions are conducted by a facilitator

(or a moderator) who is skilled at "asking questions, prompting answers and managing the flow of talk" ([33], p. 252) in order "to generate data that are 'fit-for-purpose'" ([38], p. 94).

- A computer-based **well-being questionnaire** should be developed to measure the perceived well-being of PH after interaction with PR. Annex A of the OECD guidelines for measuring subjective well-being describes items and scales for questionnaires concerned with assessing well-being (e.g., items related to happiness, comfort, satisfaction, stress, calm, or enjoyment) [29]. Ideally, the choice of the items should be decided collectively among the researchers and the staff of the nursing home where the study would take place, as in [31], with, above all, experts in well-being and even PHs being able to identify what well-being means to them, thus ensuring that the collected data can provide useful information concerning perceived well-being during interactions with PR. This requirement is important because if the usual level of well-being of an elderly person in a situation of social isolation is superior to the well-being attained by interacting with a social robot, then this situation might suggest that the social robot under study should be improved to match the expectations of end users in terms of expected benefits. The use of a questionnaire provides information concerning a posteriori self-reported well-being, which might be biased by reconstruction of the interaction in memory; therefore, focus group discussions must debate the issue of well-being through indicators of well-being that could be detected when observing the videos, not only including emotional indicators, such as enjoyment, anger, worry, unpleasantness, or pleasure, but also indicators of perceived relaxation, unease, awkwardness, tiredness, anxiety, stress, etc., which are more closely related to psychological states (see Annex A of OECD Guidelines for measuring subjective well-being [29]).
- A short, computer-based **trust, acceptability, and acceptance questionnaire** should also be developed without specific research questions: trust is also an ethical principle for AI-based systems (e.g., [28, 39]), while acceptability and/or acceptance are usually included in the evaluation of user-centered designed technologies, and the user-centered approach is recommended for the design of ethical AI-based systems [28]. Trust can be operationalized in terms of expectations of an individual with respect to the object of his or her trust [40]; for example, "I trust this social robot not to annoy me" or "I expect that this social robot is safe and will not hurt me." Therefore, the questionnaire related to trust can investigate the

dimensions of the perceived trustworthiness of social robots in terms of expected roles and benefits: this investigation requires a prior analysis of the needs and expectations of end users. According to the standard ISO 9241-210:2019[4], this requirement is a prerequisite of any user-centered designed technology. Both the acceptability and the acceptance of a technology are related to a prospective judgment concerning the introduction of this technology in the future, acceptability being measured before experience with the technology, while acceptance is measured once the individual has experienced the technology; both factors include the intention to use the technology in the future (these definitions are detailed in [41]). The items related to trust and acceptance should be expressed in a different way depending on whether the questionnaire is to be completed after PH and PR interaction or during focus group discussions: in the former, immediate feelings are measured, while the latter is a matter of rational judgment.

The main instruments are equipment for **video recording** to collect data from nonparticipant observations and **audio recording** from focus group discussions. Video recordings can be analyzed using behavior analysis software such as The Observer XT, while audio recordings and questionnaires can be analyzed with software such as NVivo [42].

3.6 Procedure

The procedure consists of two parts.

3.6.1 Data Collection from Nonparticipant Observation of Human and Robot Interactions

—**Duration**: the total duration of the study, the frequency of PR and PH meetings and their duration must, of course, be sufficient to enable the learning phase of PR, to allow participants to become familiar with each other and develop a strong rapport, and to provide time for personalized interactions to emerge. For example, the study carried out by academics from the University of Siegen (Germany) in a care home with five older adults to explore human interactions with the humanoid Pepper robot consisted of two sessions of 45–60 minutes per week over 10 weeks [30].

—**Scenarios**: PR and PH should be placed in situations featuring interactive activities, depending on the capabilities that can be implemented in the robots: conversation, game playing, physical

[4] https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-2:v1:en

The final publication is available at http://link.springer.com: Rajaonah, B., Zio, E. Social Robotics and Synthetic Ethics: A Methodological Proposal for Research. Int J of Soc Robotics (2022). https://doi.org/10.1007/s12369-022-00874-1

exercise training, cognitive stimulation, etc. In [30], three kinds of scenarios (music quizzes, exercise training and cognitive exercises for memory attention and reaction) were used, while in the study featuring the humanlike assistive communication assistive robot Matilda [31], activities included singing, dancing, reading books, playing games, phoning, walking, exercise, and dialog.

—**Technique**: The PH and PR interactions should be video recorded using the technique of overt, nonparticipant observation, but participants should be informed that they will be recorded on video. Participants should be told that the goal of the study is to study their interactions with social robots; indeed, telling them that the goal is to analyze the ethicality of these interactions may inhibit their behaviors if they think that they are being judged.

—**Questionnaires**: The well-being questionnaire should be completed by the PH at the end of each PH and PR meeting. The trust and acceptability questionnaire should be completed by the PH before starting the first day of the study, throughout the study and at its end, the goal being to obtain information concerning the dynamics of trust and acceptability in play.

3.6.2 Data Collection from Focus Group Discussions

Analysis of the video recordings and questionnaire data can provide material for the deployment of the **technique of the focus group discussion.** The material should take the form of video recorded scenarios plus a written report of the transcript of these interactions.

There should be three categories of focus group: Group G1, composed of PHs who have interacted with PR; Group G2, composed of other stakeholders; and Group G3, composed of samples from G1 and G2. Ideally, a group should be composed of four to six members [35]. Participants should be informed that these discussions will be audio recorded. They should be told that the goal of the discussion is the analysis of human and robot interactions. As mentioned in §3.1, the main goal of focus group discussions is to analyze and interpret the observed interactions, especially regarding their **ethicality operationalized in terms of socially acceptable behavior and interaction**.

The following questions could direct focus group discussions: Are expressions of PH's well-being and reluctance detectable? Which actions on the part of PR are tolerated and deemed acceptable by PH? How does PH justify his or her emotions and behaviors a posteriori? Which interactions are considered to be tolerable and acceptable by all stakeholders? Could interactions that are accepted within the context of the study be considered acceptable outside this context? None of these

questions should address morality and ethics explicitly, allowing participants to express their own opinions without being influenced.

These questions should be addressed to the three categories of focus groups (G1, G2 and G3). The first research interest is to compare the perceptions of those who have interacted with PR and those who have a "technical" view specific to their expertise to obtain knowledge concerning (i) how the issue of right and wrong behavior is apprehended by G1 during and after their interactions with PR and (ii) whether these interactions are "politically correct" as perceived by G2, as well as whether such interactions are technically, legally, medically, and socially feasible according to that group. However, overall, it is important that the interpretations of acceptable behavior and interaction are debated between these two populations, and a consensual view (i.e., interactions that are socially accepted by a majority of participants) can indeed be expected if they debate and argue their respective viewpoints. Focus groups mixing G1 and G2 are thus necessary to examine whether such debates could lead to changes in G1 and G2 members' perceived boundaries concerning acceptable PH and PR interactions and, if so, to determine the contexts in which and the arguments with which a narrowing or widening of these boundaries could be observed.

The proposed questions are merely examples of questions whose answers could provide information relevant to the research interests. The information that all stakeholders need for the design, development, deployment and use of social robots must be clarified before developing the questions for the facilitator. In addition, participants could also be invited to discuss other issues such as the following: the consequences of the introduction of social robots in the real world based on what has been observed at both the family and societal levels; the role of such robots within the family and within society; the degree of the autonomy of such robots in decision-making; the tasks that such robots will share with caregivers; the kind of guarantees and regulations that are necessary for the future deployment of social robots in nursing homes; and ways in which the social robot model discussed here can be improved for further research and future deployment (e.g., what features should a social robot have to be a useful substitute for an isolated elderly person, e.g., in terms of the capacity to detect emotional and psychological states related to well-being?).

The questionnaire concerning trust and acceptability should be completed by all participants at the end of each focus group discussion.

3.7 Data Analysis

The raw data consist of the responses provided in the completed questionnaires, the behavior protocol data provided by the video recordings of PH and PR interactions, and the verbal protocol data provided by the audio recordings of focus group discussions.

Analysis of the behavior and verbal protocol data should be carried out by using dedicated software relying on the automatic segmentation of the raw data into meaningful units and subunits according to a coding scheme that assigns a label to each unit. The coding schemes should be developed in accordance with the purpose of the study, that is, to answer the research questions and examine the hypothesis. In this context, the handbook of group interaction analysis [43] is a very useful tool that can help researchers develop or adapt their coding schemes and analyze the results.

3.7.1 Analyzing the Video Recordings

The coding scheme should focus on the interactions between humans and robots. The responses to the questionnaires concerning well-being, trust and acceptability should also be incorporated into the coding scheme. The level-1 unit of data is thus the PH-PR interaction (human and robot, humans and robot, human and robots, and humans and robots). Each video recording should be segmented into interaction level-1 units within each activity sequence. Each interaction unit should be segmented into level-2 units, the labels of which should characterize the first level: context (the scenario); actor; action; coordination of actions; emotion; coordination of emotions; goal (of emotion, action, and interaction); coordination of goals; acceptance of the other's behavior, well-being, trust and acceptability. Level-3 units should specify each level-2 unit, for instance, the type of emotion involved.

Nevertheless, these suggestions are simply indications because, to avoid wasting time and to focus only on relevant data, the coding scheme should be defined via, for instance, examination of the video recorded on the first day's study, and it should be refined when new types of interactions occur. Analysis of the occurrences over time of actions, action coordination, and expressed emotions are part of the study, and this analysis can provide information concerning the dynamics of the coconstruction of ethical behavior by PH and PR.

3.7.2 Analyzing the Audio Recordings

Similarly, the coding scheme for the analysis of the focus group discussions should be developed by examining the transcription of the first group discussion. The goal of analyzing the focus group

discussions should be to provide a kind of handbook that can help researchers find responses to their research questions (§ 3.2). The coding scheme should thus be designed in accordance with both the research questions and the questions used to direct the focus group discussions. Accordingly, the items of the questionnaire concerning trust and acceptability completed by focus group participants should also be part of the coding scheme.

For both the coding of human-robot interactions and the focus group discussions, the schemes should be developed by at least two independent researchers, as should the schemes used for analysis and interpretation of the results, to ensure research validity.

3.8 Expected Results

It must be admitted that this approach employs a difficult methodology: the setting for the deployment of the methodology is complex, the data collection procedure requires many ethical approval processes, and the data analysis requires a high degree of time and human resource consumption. However, despite these difficulties, the methodology is worth applying to deepen the understanding of our own morality, which is essential before examining the morality of artificial agents [44]. Specifically, researchers might obtain knowledge concerning the transition from human moral judgment to ethical expertise thanks to interactions with social robots in an experimental context as well as knowledge pertaining to the flexibility of moral principles in the face of new living environments with social robots. In addition, it can be expected that the results of the application of the proposed methodology will provide new insights that are useful for the design of ethical social robots and additional guidelines for designing ethical intelligent autonomous systems. Depending on the relevance of the results, for example, whether the well-being attained by interacting with PR is qualitatively and/or quantitatively inferior to the well-being experienced by the old person when he or she is isolated at home the very usefulness of social robots as such, or at least the functionalities of ethical robots that are required to provide well-being, could be put into question.

The synthetic approach is an opportunity to observe new forms of human and social robot interaction [6], but no one can predict the form that these interactions might take. Therefore, any unexpected behavior of PR and PH or any unexpected interactions between them should be analyzed carefully in the focus groups. Regarding the ethicality of PR-PH interaction, a certain degree of consensus could be expected between the people who have interacted with the robot and

those who have not regarding the acceptable aspects of observed interactions. Otherwise, it may be that the methodology is not sufficiently relevant (e.g., questions in the focus group discussions are not sufficiently in-depth) to collect useful information or that the studied model should be refined, for example, by combining their reinforcement model with an automatic recognition and expression of "emotions" as in the case of the Matilda social robot in [31].

4 Conclusions

Thanks to the discipline of synthetic ethics, researchers have the opportunity to experimentally study the development of human ethical expertise and to consider the ethicality of social robots in a new light. However, there are many other aspects that must be examined through collaboration with other disciplines prior to the deployment of such robots, including risks.

As mentioned previously, social robots evolve in a sociotechnical system in which they are an element that is interrelated with other elements, including humans, communities, organizations and institutions, technologies, and society. A project aimed at implementing social robots in society should thus assess beforehand the risks that could threaten the whole sociotechnical system: this requirement is the concept of global risk.

For our purpose, risk is defined objectively with regard to the negative consequence of an undesirable event. However, thanks to relevant standards, certain issues are likely to be overcome in the near future; for example, the IEEE Standard Association has published a standard that addresses ethical concerns during systems design, including autonomous intelligent systems (IEEE 7000™-2021[5]), and the International Organization for Standardization (ISO) is currently developing guidelines for AI applications (ISO/IEC AWI 5339) and for functional safety and AI systems (ISO/IEC AWI TR 5469)[6].

The risks that society could encounter with the introduction of social robots for the care of elderly individuals can be categorized according to the following dimensions: individual, social, societal, technological, and environmental. Examples of such risks are as follows, all discussed in [26–28, 39] with regard to ethical AI-based systems.

[5] https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html (Accessed December 1, 2021)
[6] ISO Standards for AI: https://www.iso.org/committee/6794475/x/catalogue/ (Accessed November 5, 2021)

The final publication is available at http://link.springer.com: Rajaonah, B., Zio, E. Social Robotics and Synthetic Ethics: A Methodological Proposal for Research. Int J of Soc Robotics (2022). https://doi.org/10.1007/s12369-022-00874-1

—Individual level: if the European General Data Protection Regulation[7] frames the use of digital personal data in Europe, there are remaining issues concerning the person. Namely, there are issues concerning freedom of choice regarding the free and informed consent to accept interactions with or assistance by a social robot when an elderly person suffers from dementia. In addition, assistive or companion social robots may increase the social isolation of the elderly person in terms of interpersonal relationships with humans; hence, it is important that perceived well-being with a social robot should be at least equal to perceived well-being when alone.

—Social level: abandonment of and/or irresponsibility toward elderly parents, as well as ethical issues regarding social robots with respect to the ethical behavior of humans. Furthermore, if social robots are to become humans' social partners, then the issue of their abuse may arise, as such abuse occurs among humans.

—Societal level: dehumanization of care; threats to the workforce (if substitutes in the sense of [6] become substitutes who can take the place of human caregivers); the responsibility of a robot who injures an elderly person; the responsibility of a robot who exhibits an unexpected behavior that is unethical (this issue is also a problem of technological reliability); and the fairness of the social robot regarding the cultural and religious habits of the elderly person.

—Technological level: the transparency of the ethical know-how of the social robot, especially the explainability of the transition toward behaviors and interactions that become progressively socially acceptable; cybersecurity: and the issue of avoiding situations in which a robot's ethical know-how can become unethical due a distant malicious deconditioning of the robot, which could lead to an action such that performed by Zédo.

—Environmental: environmental costs due to the mass production of social robots, the use of energy for the functioning of such robots, and the recycling of "old" robots.

To conclude, even though the proposed methodology remains an outline of such a methodology, Varela's concept of ethical know-how is a promising avenue for the design of ethical social robots, particularly in the current context of incessant machine learning progress.

Declarations

- **Funding** Not applicable

[7] https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en (Accessed November 5, 2021)

- **Conflicts of interest** Not applicable
- **Availability of data** Not applicable
- **Code availability** Not applicable

References

1. Rajaonah B, Huftier A (2020) Les robots sociaux, de la fiction à la faisabilité [Social robots, from fiction to feasibility]. In: Berrod F, Clermont P, Trentesaux D (eds) Droit et robots : Droit science-fictionnel et fictions du droit, Presses Universitaires de Valenciennes, France, pp. 51-72
2. Bates R, Goldsmith J, Berne R, Summet V, Veilleux N (2012) Science fiction in computer science education. In: Proceedings of the 43rd Technical Symposium on Computer Science Education SIGSCE'12, 161–162. https://doi.org/10.1145/2157136.2157184
3. Torras C (2010) Robbie, the pioneer robot nanny: Science fiction helps develop ethical social opinion. Interact Stud 11:269-273. https://doi.org/10.1075/is.11.2.15tor
4. Bartnek C, Forlizzi J (2004) A design-centred framework for social human-robot interaction. In: Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication RO-MAN 2004, Kurashiki, Japan. https://doi.org/10.1109/ROMAN.2004.1374827
5. Shaw-Garlock G (2009) Looking forward to sociable robots. Int J of Soc Robotics 1: 249-260. https://doi.org/10.1007/s12369-009-0021-7
6. Dumouchel P, Damiano L (2017) Living with robots. Harvard University Press, Massachusetts
7. Ellemers N, van den Bos K (2012) Morality in groups: On the social-regulatory function of wright and wrong. Soc Pers Psychol Compass 6:878-889. https://doi.org/10.1111/spc3.12001
8. Varela FJ (1999) Ethical know-how. Action, wisdom, and cognition. Standford University Press, California
9. Varela F (2004) Quel savoir pour l'éthique ? Action, sagesse et cognition [Ethical know-how. Action, wisdom, and cognition]. Éditions La Découverte, Paris
10. Wallach W (2008) Implementing moral decision-making faculties in computers and robots. AI & Soc 22:463–475. https://doi.org/10.1007/s00146-007-0093-6
11. Dawson M R W (2002) From embodied cognitive science to synthetic psychology. In: Proceedings of the First IEEE International Conference on Cognitive Informatics. IEEE, New York. https://doi.org/10.1109/COGINF.2002.1039276
12. Damiano L, Cañamero L (2010) Constructing emotions: Epistemological groundings and applications in robotics for a synthetic approach to emotions. In: Proceedings of the 36th Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB'10), Leicester, United Kingdom
13. Seibt J (2016) 'Integrative Social Robotics' - A new method paradigm to solve the description problem and the regulation problem? In: Seibt J, Nørskov M, Andersen S S (eds.), Robophilosophy 2016 / TRANSOR 2016, vol. 290. What social robots can and should do. IOS Press, The Netherlands, pp 104-115. https://doi.org/10.3233/978-1-61499-708-5-104
14. Damiano L (2009) Creative coordinations: Theory and style of knowledge in P. Dumouchel's emotions. World Futures 65:568–575. https://doi.org/10.1080/02604020903300568

15. Heylighen F (1989) Self-organization, emergence and the architecture of complexity. In: Proceedings of the 1st European Conference on System Science, Paris), 23-32. http://pespmc1.vub.ac.be/Papers/SelfArchCom.pdf
16. Hollnagel E, Woods D (2005) Joint cognitive systems: Foundations of cognitive systems engineering. CRC Press, Boca Raton, Florida
17. Varela F, Thompson E, Rosch E (1993) L'inscription corporelle de l'esprit. Sciences cognitives et expérience humaine. Éditions du Seuil, Paris
18. Varela FJ, Thompson E, Rosch E (2016) The embodied mind. Cognitive Science and Human Experience, revised edition. The MIT Press, Massachussetts
19. Svensson H, Ziemke T (2004) Making sense of embodiment: Simulation theories and the sharing of neural circuitry between sensorimotor and cognitive processes. In: Proceedings of the 26th Annual Conference of the Cognitive Science Society, Chicago, Illinois, pp 1309–1314
20. Rizzolatti G, Fabbri-Destro M (2008) The mirror system and its role in social cognition. Curr Opin Neurobiol 18:179–184. https://doi.org/10.1016/j.conb.2008.08.001
21. Cattaneo L, Rizzolatti G (2009) The mirror neuron system. Arch Neurol 66:557–560. https://psycnet.apa.org/doi/10.1001/archneurol.2009.41
22. Hoffman G (2012) Embodied cognition for autonomous interactive robots. Top Cogn Sc 4:759–772. https://doi.org/10.1111/j.1756-8765.2012.01218.x
23. Breazeal C, Gray J, Berlin M (2009) An embodied cognition approach to mindreading skills for socially intelligent robots. Int J Rob Res 28:656-680. https://doi.org/10.1177%2F0278364909102796
24. Balkenius C, Cañamero L, Pärnamets P, Johansson B, Butz MV, Olsson A (2016) Outline of a sensory-motor perspective on intrinsically moral agents. Adapt Behav 24:306–319. https://doi.org/10.1177%2F1059712316667203
25. Winkle K, Caleb-Solly P, Turton A, Bremner P (2020) Mutual shaping in the design of socially assistive robots: A case study on social robots for therapy. Int J Soc Robot 12:847–866. https://doi.org/10.1007/s12369-019-00536-9
26. Dilhac MA, Abrassart C, Voarino N (2018) Montréal Declaration for a responsible development of artificial intelligence. Université de Montréal, Montréal, QC
27. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019) Ethically Aligned Design. First Edition. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. IEEE. https://ethicsinaction.ieee.org/. Accessed February 21, 2021
28. AI HLEG (2019) Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence, European Commission, Brussels. https://ec.europa.eu/futurium/en/ai-alliance-consultation. Accessed February 21, 2021
29. OECD (2013) Guidelines on measuring subjective well-being. OECD Publishing, Paris. https://doi.org/10.1787/9789264191655-en
30. Carros F, Johanna M, Löffer D, Unbehaun D, Matthies S et al (2020) Exploring human-robot interaction with the elderly: Results from a ten-week case study in a care home. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, New York. https://doi.org/10.1145/3313831.3376402
31. Khosla R, Chu MT, Nguyen K (2013) Enhancing emotional well-being of elderly using assistive social robots in Australia. In: Proceedings of the 2013 International Conference on Biometrics

and Kansei Engineering (ICBAKE), Tokyo, Japan, 41-46. https://doi.org/10.1109/ICBAKE.2013.9

32. Strasser A (2020) Social norms for artificial systems. In: Nørskov M, Sibt J, Quick O S (eds) Culturally Sustainable Social Robotics - Proceedings of Robophilosophy 2020, 295-304. https://doi.org/10.3233/FAIA200926
33. Scott J (ed) (2014) A dictionary of sociology, 4th edn. Oxford University Press, Oxford
34. Gulliksen J, Göransson B, Boivie I, Blomkvist S, Persson J, Cajander Å (2003) Key principles for user-centred systems design. Behav Inf Techno 22:397-409. https://doi.org/10.1080/01449290310001624329
35. Kitzinger J (1994) The methodology of focus groups: The importance of interaction between research participants. Sociol Health Illn 16:103-121. https://doi.org/10.1111/1467-9566.ep11347023
36. Ravitch SM, Carl N M (2020) Qualitative research: Bridging the conceptual, theoretical, and methodological, 2nd edn. SAGE Publications, Inc, California
37. Plummer-D'Amato P (2008) Focus group methodology. Part 1: Considerations for design. Int J Ther Rehabil 15:69-73. https://doi.org/10.12968/ijtr.2008.15.2.28189
38. Barbour R (2018) Doing focus groups. SAGE Publications Ltd, London
39. Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics. Nat Mach Intell 1:389-399. https://doi.org/10.1038/s42256-019-0088-2
40. Hardin R (2001) Conceptions and explanations of trust. In: Cook KS (ed) Trust in society. Russel Sage Foundation, New York, pp 3-39
41. Adell E, Várheli A, Nilsson (2014) The definition of acceptance and acceptability. In: Regan MA, Horberry T, Stevens A (eds) Driver acceptance of new technology. Theory, measurement and optimization, 1st edn. Taylor & Francis Group. https://doi.org/10.1201/9781315578132
42. Glüer M (2018) Software for coding and analyzing interaction processes. In Brauner E, Boos M, Kolbe M (eds.) The Cambridge handbook of group interaction analysis. Cambridge University Press, United Kingdom, pp 245–273
43. Brauner E, Boos M, Kolbe M (eds) (2018) The Cambridge handbook of group interaction analysis. Cambridge University Press, United Kingdom
44. Hunyadi M (2019). Artificial Moral Agents. Really? In: Laumond JP, Danblon E, Pieters C (eds), Wording robotics. Discourses and representations on robotics. Springer, Switzerland, pp 59–69. https://doi.org/10.1007/978-3-030-17974-8_5