International Journal of Social Robotics (2023) 15:1977–1980
https://doi.org/10.1007/s12369-023-01084-z

Experimental and Integrative Approaches to Robo-ethics. An Introduction

Francesco Bianchini[1] · Luisa Damiano[2] · Edoardo Datteri[3] · Pierluigi Graziani[4]

Published online: 4 December 2023

The development and diffusion of social robots are raising several ethical and legal concerns. Traditional approaches to robo-ethics have relied in the past primarily on philosophical analysis to address these issues. However, there is a growing recognition that experimental and integrative approaches are also needed.

To promote studies in this direction in 2019, a workshop was organized at the International Conference on Social Robotics in Madrid. The main goal of the workshop titled "Experimental and Integrative Approaches to Robo-ethics" (EIAR 2019) was to discuss the potentialities, limits, and methods of this new direction, with a specific focus on the role that experiments can play in addressing ethical and legal issues concerning (social) Human-Robot Interaction (HRI). The workshop presentations included several concrete pioneering experimental and integrative approaches in robo-ethics, which instigated structured discussions about their further development, improvement, and diffusion. The ultimate goal was to initiate the formation of an interdisciplinary research community engaged in the development of a well-defined research line in experimental and integrative robo-ethics.

This volume brings together some of the papers presented during the workshop with papers that were selected through a peer-reviewed international call for papers via this journal. The presented work covers a wide range of topics, including:

- the dilemma of control in social robotics;
- perceived safety in human-robot interaction;
- artificial emotions in a human-robot interaction;
- democratization of robot technology;
- effective interdisciplinary exchange in the development of ethical regulation of robotics;
- ethical design methodologies in social robotics;
- types of artificial social agents and related ethical issues in human-technology social interaction;
- psychological mechanisms involved in human-robot social interaction;
- persuasive social robots.

These very different problem spaces highlight the importance of interdisciplinary collaboration in the field of robo-ethics.
Some additional details about the articles in the volume can be found below.

The first article addresses the problem of the Collingridge Dilemma or dilemma of control, according to which once a technology is produced, it is not possible to establish all its consequences in advance. In the event that complex problems arise due to this technology and it is already sufficiently widespread, this diffusion would make it difficult to implement possible solutions. This dilemma is addressed in the field of social robotics where technological progress has a significant impact not only from a technical point of view, but also from a social, regulatory and political point perspective. To overcome this dilemma, various strategies have been implemented which call into question anticipatory aspects, i.e. hypotheses of predictions concerning the possible future scenarios with the technologies under observation; or socio-technical aspects, which do not include all involved stakeholders; or finally aspects based not only on interaction but on the actual socialization created and implemented by these technologies. This latter approach suffers from a high degree of uncertainty (even more than the two previous cases). The proposal of the author is to arrive at

✉ Luisa Damiano
luisa.damiano@gmail.com

[1] University of Bologna, Bologna, Italy

[2] IULM University of Milan, Milan, Italy

[3] University of Milano-Bicocca, Milan, Italy

[4] University of Urbino, Urbino, Italy

the inclusion of ethical aspects in these types of analysis to promote a responsible approach to social robotics. To this end, an inclusive approach that considers all affected stakeholders involved with these technologies is considered relevant. This inclusiveness cannot ignore transdisciplinary considerations, both in terms of the involved stakeholders and of the values constituting the background of the conceptual structure within which not only a technology is built and operates in the first instance, but also in a derivative way; as a producer of contents and operational social practices. To this principle the related principle of responsibility in designing, building and monitoring robotic technologies used in the social field is added, starting not only from scientific-technological conceptual frameworks, but also from social values in a pluralist perspective. This layout, which is difficult to achieve in many different scientific enterprises, is no less arduous in the creation of social robotics artifacts, but seems even more compelling and perhaps connected to an area of more evident and direct implications and analyses from a methodologically integrative perspective.

The second article analyzes the problem of safety in social robotics from the specific standpoint of perceived safety by users. The perception of safety is added to physical safety, and to concerns about possible cyber-intrusions by agents capable of using social robotic artifacts for malicious purposes like data theft. Perceived safety includes psychological, moral and social aspects. The authors trace the various proposals present in the literature to six principles, among which reliability, the sense of control and the predictability of artifacts' behavior stand out. They add to these principles a series of contextual factors that cannot be defined according to an exhaustive list, but can however be tested frequently in the field. The context of use of social robotic artifacts can greatly influence the perception of safety in interaction with social robots, depending on the purposes, attitudes and intentions attributed to such robots. The tools to evaluate the perception of safety in relation to interaction with social robots are not yet well developed. Generally, they include reports by human agents operating in a controlled environment in the presence of social robots, or surveys aimed at understanding the psychological and emotional aspects of human agents. These tools are unbalanced on the human aspects of interaction and seem to address less those related to the robotic artifacts themselves. The authors, in proposing a taxonomy for perceived safety in HRI, point out that the relationship between human-related factors and robot-related factors is complex. Depending on where the emphasis is placed, reports and measurements can vary, as the perception of safety is also prone to the influence of human-related social factors. The experiment proposed by the authors leads to the rather interesting conclusion that the aspects of perception of unsafety are more quantifiable than those linked to safety. This can be a good starting point for structuring future research in an integrative perspective combining operational practices with conceptual elements that influence the behavior of the robotic artifact not only in the design phase, but also in the activity in open-ended environments.

In ***the third article*** the problem of integrating ethical approaches in robotics and AI is addressed based on the discussion of an interdisciplinary project dedicated to "democratize collaborative robot technology". This article reports on the author's direct engagement in this project. Following an overview of existing perspectives on robo-ethics, the author focuses on the collaborative robot (or cobots) project at the center of the article, and examines the activities supporting of the project's ambition of "democratizing" this technology with a particular attention to related challenges and opportunities. On these bases, the article promotes, as key ingredients for the future of successful integrative ethics approaches, interdisciplinary methodologies, engagement in criticism, and new, specific modes of anticipation. With regard to future research, the author prospects factors that can generate useful insights, including the discussion on the integration of ethics, which can produce new ethical frameworks, and the expansion of research in new, real-world contexts of interaction between people and robots, where integrative ethical approaches have great potential to provide larger frameworks, able to explore related socio-ethical implications.

The fourth article presents an interdisciplinary approach tackling the general problem of establishing a fruitful dialogue between different actors, belonging to different disciplines, that can offer relevant insights for the elaboration of effective ethical regulation of robotics. At the center of the research work presented in this article is the proposal submitted by Madi Delvaux to the European Parliament, on May 31st 2016, with the twofold goal of establishing regulations, concerning the civil robotics sector, for the whole European Union. After a detailed description of this motion, usually called Draft Report on Civil Law Rules on Robotics, the authors present, analyze and discuss the opinions that roboticists, belonging to different robotics labs based in Europe, expressed with regard to it by completing a questionnaire in the context of dedicated seminars. The research work presented in the article shows roboticists' interest in the ethical dimension of their work and related policy, and, at the same time, reveals their belief that there are significant misunderstandings in how policy makers view robotics and AI. Considering roboticists' lack of trust in the role that experts outside the field can play in regulating robotics effectively, the authors propose an integrative approach to robo-ethics as a way to break down these misunderstandings.

Article five explores the impact of social robots on the relationship between humans and robots from the point of view of emotions. Internalist positions regarding artificial emotions are hard to maintain and imply the recognition of mechanisms that would implement the possibility of having emotions (and therefore their perception) by robots. This problem has significant theoretical obstacles concerning what a robot's self-perception or self-awareness about emotions might be. From a social point of view, however, emotions are investigated for aspects related to bodily aspects and behavior, as well as facial expressions. This is the perspective adopted by the authors of the article. Social robotics lends itself well to such reflections, because the focus of interest lies in the interaction between human beings and robotic artifacts, and therefore in the relationship they establish and which conveys an emotional sense generated by the interaction itself. Anthropomorphic robots or robots with anthropomorphic behavior are the most suitable for generating emotional responses. These give rise to quite a few ethical problems, starting with what the authors consider to be among the most relevant: the problem of attributing reliability. The predominant aspect of the relational dynamics lies in the robot's understanding of human's emotions and in the robot's appropriate emotional communicative response. The authors discuss systems capable of providing these emotional responses, but above all capable of understanding the emotional aspects of human beings. They are therefore robotic artifacts which generate the problem of their reliability, but also pose the problem of the human being transparency as regards the emotions communicated, and therefore the more general question about the type of ethical treatment that this kind of interactions require/may require in the future developments and applications, a request that cannot ignore an integrative approach. The human-centric perspective shows its fragility, and therefore the need for adequate ethical and ethical-synthetic treatments, in dependence on the increasing anthropomorphic robotic context which we act in.

Article six discusses ethical issues related to the field of Socially Assistive Robotics (SAR), focusing specifically on educational settings in underfunded contexts as a frontier application domain. The authors present empirical research dedicated to determine and discuss ethical challenges, and associated pedagogical issues, related to the application of SAR in the above-mentioned settings and contexts. More specifically, the research work presented in this article consists of a 5-week "in-the-wild" user study, procedurally based on video recordings and interaction analyses, realized with 12 kindergarten children attending a community school with low funding levels in New Delhi (India). As the authors emphasize, their study – detailed in the article in all its aspects – allowed them to individuate four fundamental ethical aspects to be considered in the context of SAR projects targeting the use of social robots in communities with low levels of education funding. Schematically, these ethical challenges are presented as the following series of potential obstacles: (a) accent and language issues which could impede pedagogical activities; (b) harmful technology malfunctions; (c) trust and deceit issues; (d) issues related to ecological sustainability.

Article seven addresses the topic of experimental and integrative approaches to robo-ethics through the definition of a methodology directed to support ethical design of social robots for vulnerable individuals. At the theoretical level, the authors ground this methodology in three key elements: (a) Francisco Varela's embodied notion of "ethical know-how", of which the authors offer an original development in the field of social robotics; (b) the "synthetic" approach to robo-ethics introduced by Luisa Damiano and Paul Dumouchel for the development of an effective experimental and transdisciplinary ethical inquiry on the interactions between humans and social robots; (c) a model of intrinsically moral robots developed by Christian Balkenius and colleagues. At the procedural level, the authors root the methodology proposed in the article in a combination of non-participant observation, focus group discussion and questionnaires. Based on these theoretical and procedural components, the authors propose their methodology as an effective tool to gather information on how the ethical dimension of human-robot interactions, conceptualized as a growing ethical expertise on both sides, can be significantly improved.

The eighth article develops the topic of robo-ethics based on a philosophical exploration of the embodiment of natural and artificial agents, and its role in determining the features of human-technology interaction. The author engages in the comparison of two types of artificial social agents, whose differences derive from divergent forms of embodiment, understood as a specific way of being present in the world. According to the taxonomy proposed by the author in this article, the first type of artificial social agents consists of "true robots", while the second of "analytic agents". The article describes *true robots* as three-dimensional physical machines, characterized by individuality, possibilities for environmental manipulation and mobility in the physical space, and conceptualizes *analytic agents* as entities that, similar to mobile phone applications, can act in social environments only when integrated in complex systems that are composed of a variety of technologies. Based on an analysis of the differences between these two types of social agents, and of the elements that distinguish them from human agents, the article delineates their different ways of interacting with humans, and elaborates on the ethical and political dimensions of these interactions.

Article nine presents a theoretical framework to understand the psychological mechanisms in Human-Companion Robot Interactions, focusing on Sexual Robotics as a case study. It distinguishes Sexual Robots from sex toys, highlighting collusive dynamics and potential consequences like paraphilic fixation and user infantilization. The discussion extends to Companion Robotics in general, exploring if relational dynamics in Human-Robot Interaction (HRI) can shift to human relations. The global interest in Social Robots, especially in healthcare, is discussed, emphasizing the accelerated integration of Companion Robots during the COVID-19 pandemic. The paper delves into the debate on Sexual Robots, contrasting polarized views on their societal impact. The symbolic consequences argument is examined, proposing a parallelism with virtual reality and augmented reality in the context of embodied cognition. The authors argue that collusive dynamics observed in Sexual Robots may apply to Companion Robotics in general, impacting users' relational abilities. An experimental setup using the humanoid robot NAO is proposed to test the impact of collusive interactions on users' ability to manage relational frustration. The paper concludes by discussing both the potential therapeutic use of these artifacts within a clinical setting and the limitations of the paper itself due to an androcentric perspective and a lack of analysis of the fetishization of Sexual Robots. The authors call for further research and experimental data to validate their theoretical framework and hypotheses.

Article ten deals with Persuasive Technologies, a Human-Computer Interaction (HCI) field focusing on designing systems that influence user attitudes and behaviors. Creating persuasive robots, especially in Social Robotics, presents a research challenge. The paper proposes a model integrating cognitive aspects, storytelling, ethics, and rhetorical techniques within the ACT-R cognitive architecture. The robot employs various persuasive techniques to discuss COVID-19 rules and vaccines, considering the interlocutor's awareness and ethical considerations. The paper's structure includes an overview of related works, the agent model integrated into ACT-R, implementation details, an interaction example, and evaluation results. The COVID-19 topic, widely debated, serves as a test case, reflecting changing public opinions and behaviors. The preliminary experimentation identifies the effectiveness of storytelling, emphasizing ethical principles for persuasive strength. Classical rhetoric techniques, such as *ad verecundiam* and framing, show persuasive effects, while *ad populum* does not. The system targets users opposing vaccines or COVID-19 rules, revealing varying agreement levels. Future work aims to implement the model in a physical social robotic platform to leverage embodied features. Real-world engagement may overcome the limitations of virtual avatars, particularly in handling skeptical users. The study emphasizes the importance of an ethical framework for AI systems, ensuring the responsible use of persuasive techniques. The paper acknowledges the need for further investigation into the ethical acceptability of these techniques and suggests an "ethics by design" approach.

The authors of these articles use various methods and approaches to explore the ethical challenges social robots pose. Some articles use experimental methods to study how people interact with social robots. Others use philosophical analysis to develop ethical guidelines for social robots. Still, others use an integrative approach to combine experimental and philosophical methods.

This volume is thought to be a valuable resource for researchers, practitioners, and anyone interested in the ethical implications of social robots. It provides a comprehensive overview of the field and offers new insights into the challenges and opportunities of developing ethical guidelines for social robots.

Future research in social robotics will probably have to deal increasingly with integrative approaches, both because this technology will be more and more widespread, with all the prediction limits that an interactive technology in the real world brings with it, beyond the fact that it is robotics, and because robotics in its social meaning seems to bring into play peculiarly the close interrelationship between experimental data and theoretical aspects epistemologically characterizing the most complex scientific fields. Therefore, the integrative approaches to this discipline are part of an interdisciplinary and transdisciplinary line of research that increasingly contributes to scientific advancement and to overcoming the fragmentations and specializations of scientific languages and methods. Not by chance, the social we are talking about is not the collective one, but the interactive one.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.