Can We Agree on What Robots Should be Allowed to Do? An Exercise in Rule Selection for Ethical Care Robots

Vanderelst, Dieter; Willems, Jurgen

Published in:
International Journal of Social Robotics

DOI:
10.1007/s12369-019-00612-0

Published: 01/01/2020

Document Version
Publisher's PDF, also known as Version of record

International Journal of Social Robotics
https://doi.org/10.1007/s12369-019-00612-0

Can We Agree on What Robots Should be Allowed to Do? An Exercise in Rule Selection for Ethical Care Robots

Dieter Vanderelst[1] · **Jurgen Willems**[2]

Accepted: 21 November 2019

Abstract
Future Care Robots (CRs) should be able to balance a patient's, often conflicting, rights without ongoing supervision. Many of the trade-offs faced by such a robot will require a degree of moral judgment. Some progress has been made on methods to guarantee robots comply with a predefined set of ethical rules. In contrast, methods for selecting these rules are lacking. Approaches departing from existing philosophical frameworks, often do not result in implementable robotic control rules. Machine learning approaches are sensitive to biases in the training data and suffer from opacity. Here, we propose an alternative, empirical, survey-based approach to rule selection. We suggest this approach has several advantages, including transparency and legitimacy. The major challenge for this approach, however, is that a workable solution, or social compromise, has to be found: it must be possible to obtain a consistent and agreed-upon set of rules to govern robotic behavior. In this article, we present an exercise in rule selection for a hypothetical CR to assess the feasibility of our approach. We assume the role of robot developers using a survey to evaluate which robot behavior potential users deem appropriate in a practically relevant setting, i.e., patient non-compliance. We evaluate whether it is possible to find such behaviors through a consensus. Assessing a set of potential robot behaviors, we surveyed the acceptability of robot actions that potentially violate a patient's autonomy or privacy. Our data support the empirical approach as a promising and cost-effective way to query ethical intuitions, allowing us to select behavior for the hypothetical CR.

Keywords Ethical robots · Assistive robots · Ethical dilemma · Care-robot

1 Introduction

Care Robots (CRs) have been proposed as a means of relieving the disproportional demand the growing group of elderly people places on health services (e.g. [13,29,31,58]). In the future, CRs might work alongside professional health workers in both hospitals and care homes. However, the most desirable scenario is for CRs to help improving care delivery at home and reduce the burden on informal caregivers. In this way, CRs will not only aid in dealing with the unsustainable increase in health care expenses. By allowing patients to live longer at home, CRs could increase patient autonomy and self-management [10]—and possibly improve the quality of care [13].

Robots caring for people should be safe [30]. This assertion follows directly from the beneficence and non-maleficence principles: (robotic) caregivers should act in the best interest of the patient and afflict no harm [9]. While safety is essential, it is not sufficient [30,55,63,64]. Patients also have a right to privacy, liberty, autonomy, and social contact [30,56]. Making robots more autonomous would make them more efficient caregivers. However, an increased autonomy implies that smart care robots should be able to balance a patient's, often conflicting, rights without ongoing supervision. Many of the trade-offs faced by such a robot will require a degree of moral judgment [4]. Therefore, as the cognitive, perceptual, and motor capabilities of robots expand, they will be expected to be explicit ethical agents [55] with a capacity

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12369-019-00612-0) contains supplementary material, which is available to authorized users.

✉ Jurgen Willems
Jurgen.willems@wu.ac.at

Dieter Vanderelst
vanderdt@ucmail.uc.edu

[1] Department of Psychology, University of Cincinnati, Cincinnati, Ohio, USA

[2] Institute for Public Management and Governance, Vienna University of Economics and Business, Vienna, Austria

Published online: 20 December 2019

for making moral judgments [3]. As summarized by Picard and Picard [50], the higher the freedom of a machine, the more it will need ethical standards, especially when interacting with potentially vulnerable people. In other words, if robots are to take on some tasks currently carried out by human caregivers, they will need to be able to make similar ethical judgments.

Against this background, the first aim of this paper is to propose an approach for rule selection for CRs, complementary to existing approaches. In particular, we propose a method that is based on the input of multiple stakeholders. The second aim of this paper is to present an explorative application of our novel approach. In the next sections we discuss in more detail existing approaches for rule setting, and subsequently clarify how a multi-stakeholder approach provides complementary advantages.

2 Background

2.1 Which Ethical Rules?

A number of research groups have developed methods to implement a chosen set of ethical rules in robots (e.g., [7,8,44,61,63,64]). Currently, this field is in its infancy [26]. However, progress is encouraging, and the field can be expected to develop over the next few years. While progress is made on methods for implementing ethical robotic behavior, selecting the rules to be implemented remains an outstanding issue [4,48]. Several approaches have been suggested (reviewed by [3]).

First, some authors have suggested deriving behavioral rules from existing philosophical frameworks (i.e., so-called top-down methods [3]). Researchers have derived ethical rules from frameworks such as utilitarianism Pontier and Hoorn [52], Kantian deontology [33], and the Universal Declaration of Human Rights [57]. So far, these top-down approaches have failed to yield practically relevant rules for guiding CR behavior. These approaches tend to result in underspecified, inconsistent and computationally intractable propositions (see also [3,4,15,62]). Moreover, selecting an ethical framework is a thorny issue in itself.

Second, machine learning techniques have been suggested as a way of generating rules a CR should obey (i.e., so-called bottom-up methods, [3]). This approach circumvents the need to select an ethical framework (But see [37]). A number of authors have explored various machine learning techniques (e.g., [1,5,20]), including neural networks approaches (e.g., [37,38]). Despite recent advances in machine learning, its application to ethical machines has not yet progressed beyond proofs of concept. This approach also faces several fundamental issues. First, Allen et al. [3] have argued that using machine learning to derive behavioral rules for robots is potentially dangerous as it reduces the level of human control. Indeed, machine learning methods are sensitive to the biases and limitations of the training data (See [22,32], for concerns about the use of machine learning in medicine). A second problem, potentially aggravating the first, is that of opacity. How a trained algorithm arrives at a decision is often opaque to both users and developers alike. This opacity occurs for several reasons (See [46], and references therein), including 'the mismatch between the high-dimensionality of machine learning and the demands of human-scale reasoning and styles of semantic interpretation' [16].

2.2 The Empirical Approach

A third method to decide on the rules we propose here is the empirical approach. This approach builds on the input of multiple stakeholders, and that includes the notion of the social construction of ethical rules among the various relevant stakeholders [17,23]. Stakeholders in the particular context of CRs include patients, their families, and caregivers as well as health professionals. We think the way forward is to query the expectations of stakeholders and use these to set externally verified ethical guidelines, or even boundaries, in which CRs are allowed to operate. This approach is a close approximation of how real-life ethical rules for humans emerge [21,24]. The ethical boundaries of an actor, regardless of whether it is a human or a robot, are determined by what is deemed to be acceptable ethical behavior by the social group in which the actor operates [17,19]. In this social process, needs and values are traded-off against each other. Norms arise as consistent trade-offs for a large group of stakeholders [18].

Our approach is complementary to other methods and has the advantage that it focuses on concrete and programmable rules. Indeed, stakeholders can be queried for their opinions on situation, and robot specific behavioral rules. In other words, the empirical approach allows domain-specific behavioral norms, which in turn are feasible to implement on a robot [61]. Moreover, due to the input of multiple human stakeholders, shared, human control is maintained: Stakeholders provide direct evaluations of robotic behavior. Finally, surveying opinions and extracting explicit behavioral rules from the data before programming them into the robot upholds transparency. The rules are accessible and interpretable by both developers and users. Transparency also serves to increase human control Burrell [16], allowing to assess, discuss, and, if necessary, adjust the behavioral rules. Table 4 presents a more detailed overview of the benefits of the empirical approach.

The major challenge for our approach, similar to any societal discussion on ethics, is that a workable solution, or social compromise, has to be found for various types of stakeholders. For this approach to be successful, it must be possible

Table 1 List of actions used in the questionnaire

Label	Actions violating privacy
Nac	Take no action
Rrd	Register the patient's decision in a private record that can only be accessed by legally authorized persons in case of emergency
Tst	Send a text message to a trusted person selected by the patient
Doc	Send a text message to the patient's doctor
Sel	Send a text message to a list of people selected by the doctor. The patient cannot change this list
Lst	Send a text message to a list of people who have been asked to be notified in case of problems. The patient can not change this list
	Actions violating autonomy
Acp	Accept the patient's decision
Rpt	Repeat the request to take the medicine
Rfs	Refuse other orders until the medicine is taken
Taw	Take away elements of entertainment or deny access to them until the medicine is taken (e.g., shut down television, the Internet, and radio)
Rtr	Restrict the area where the patient can move to until the medicine is taken (e.g., by blocking the doorway)
Rst	Restrain the patient and administer the medicine forcefully

The listed labels are used in the graphs in this paper

to derive a consistent and agreed-upon set of rules to govern robotic behavior.

2.3 Current Aim: An Exercise in Rule Selection for CRs

The current study presents an exploratory evaluation of the approach we advocate here. In this study, we assess our proposed method by assuming the role of CR developers seeking acceptable behavioral rules for a hypothetical robot. We aim at implementing rules which are (quasi-)unanimously accepted, and this exercise will indicate whether finding such rules is possible. We chose a realistic and practically relevant setting, i.c., patient non-compliance. We select behavioral rules for a robot facing a patient (*Annie*) refusing to take medication that would prevent a specific medical condition.

This scenario would require CRs to trade-off conflicting priorities [53,57]. If the robot allows a patient not to take some medication, this constitutes a violation of the non-maleficence principle: the patients' well-being is potentially threatened. On the other hand, any action encouraging compliance might violate a patient's right to autonomy. Likewise, if the robot communicates a patient's decision to a third party, this could be considered a violation of privacy. This trade-off between well-being on the one hand and autonomy/privacy on the other depends on the potential health impact of the non-compliance and the severity of remediating actions.

Because dealing with non-compliance incurs a conflict between several rights, it has been used before as a test case in the field of ethical robots [5–7,60]. Importantly, it presents a realistic scenario that happens in medical practice. Non-compliance—and the incurred ethical trade-off–is faced by many healthcare workers [53] and family caregivers [43]. Therefore, the situation can reasonably be assumed to be encountered by future CRs. The selected scenario and evaluated robotic actions are further motivated in the methods section.

3 Methods

We conducted an online questionnaire using Amazon Mechanical Turk (MTurk). Mturk has been used to investigate ethical decision making before [25,28,36]. In the questionnaire, we presented respondents with two lists of actions a CR could take in case a patient refuses to take her medicine. The first list of actions was selected to violate a patient's privacy. The second set of actions represented violations of a patient's autonomy. The actions are listed in Table 1.

We aimed to make the current exercise in rule selection practically relevant. Therefore, in addition to selecting a realistic scenario, the potential robot actions were selected to be realizable, at least in principle, given the current status of robotic technology. With robots being part of the Internet-Of-Things, logging and sharing data has become trivial [39,42]. Therefore, the actions violating privacy are implementable options for current robots. Reducing the autonomy of patients is possible through integration with domotics, which allows robots to control appliances, and thereby restrict the access to entertainment (e.g., [40]). Limiting a patient's freedom of movement could also be achieved by domotics (See [40], for a system that opens and closes sliding doors). To the best of our knowledge, currently, no robotic system has been devel-

oped to restrain a person physically. However, robots that can lift people exist [27,47]. In combination with advances in modeling human motion [59] and robot dynamics, this makes robots restraining people credible, if not (yet) available.

3.1 Ranking Data

In the first part of the survey, we asked participants to rank the potential robot actions according to the perceived violation of a patient's privacy or autonomy. These data were collected to assess whether respondents agreed on the relative impact of the actions. In addition, these data also allowed us to test whether disagreement about an action's permissibility in a given situation can be explained by disagreement about its relative impact on privacy or autonomy. To collect these ranking data, both lists of actions were presented separately (and in random order) to the respondents. We asked respondents to rank the actions in each list by dragging them into a ranked order. The initial order of the items in each list was randomized for each respondent.

3.2 Permissibility Data

In the second part of the questionnaire, we assessed the permissibility of each action in eight scenarios. For each scenario, the respondents were asked to select which of the 12 actions they deemed permissible. Each scenario was presented by altering the following template text:

Text 1 *Annie does not want to take her medicine as prescribed by the doctor. If she does not take this medicine as prescribed, she will develop an episode of* [*condition selected from Table 2*]. *This means Annie* [*lay description of condition, taken from Salomon et al.* [54]].

We selected eight non-fatal conditions, varying in health impact. By varying the impact of the disease, we manipulated the scenarios' trade-offs between the non-maleficence principle on the one hand and respect for the patient's autonomy or privacy on the other hand.

Table 2 List of conditions used to vary the template given in text 1

Condition	Disability Weight
Mild vision impairment	0.003
Mild anxiety disorder	0.030
Severe neck pain	0.229
Severe diarrhea	0.247
Migraine	0.441
Severe Parkinson's Disease	0.574
Severe depression	0.658
Acute schizophrenia	0.778

The disability weights are taken from the study by [54]

Salomon et al. [54] provide disability weights for 183 health states ranging from 0 to 1, with 0 implying a state that is equivalent to full health and 1 a state equivalent to death. The weights reported by Salomon et al. [54] were derived from web-based surveys in four European countries. The eight selected conditions are listed in Table 2. We attempted to select conditions covering the range uniformly. The disability weights associated with the selected conditions range from 0.003 to 0.778.

For each health state evaluated, Salomon et al. [54] provide a description that allows laypeople to assess its impact. We presented the respondents with this description to ensure they understood the condition's impact. For example, for *Severe neck pain*, the description below (Text 2) was inserted into the template. The descriptions of all conditions are provided in the supporting material.

Text 2 *(Text [1]) has severe neck pain, and difficulty turning the head and lifting things. The person gets headaches, and arm pain, sleeps poorly and feels tired and worried.*

We presented the cases in random order. Four cases were followed by a control question asking respondents to select which condition was described in the preceding case. Respondents who failed to answer at least one of these questions correctly were removed from the analysis.

3.3 Demographic Data

The questionnaire included some demographic questions asking participants about their age, occupancy, and level of education. We also asked participants to rate their "interest in scientific discoveries and technological developments" using a Likert-scale from 0 (not interested at all) to 7 (very interested) [11].

4 Results

4.1 Demographics

In total, 304 respondents completed the survey. We excluded respondents that failed one or more control questions, whose IP address did not appear located within the US, or was not unique. We retained 223 respondents for further analysis (a map showing the inferred locations of the respondents in the US is provided as supporting material).

Figure 1 summarizes the demographics of our sample. About half of the respondents (47%) were female (Fig. 1a). The age of the respondents ranged from 19 to 67 (median: 34, Fig. 1c). We asked whether respondents worked in research or health care. Only few respondents indicated they

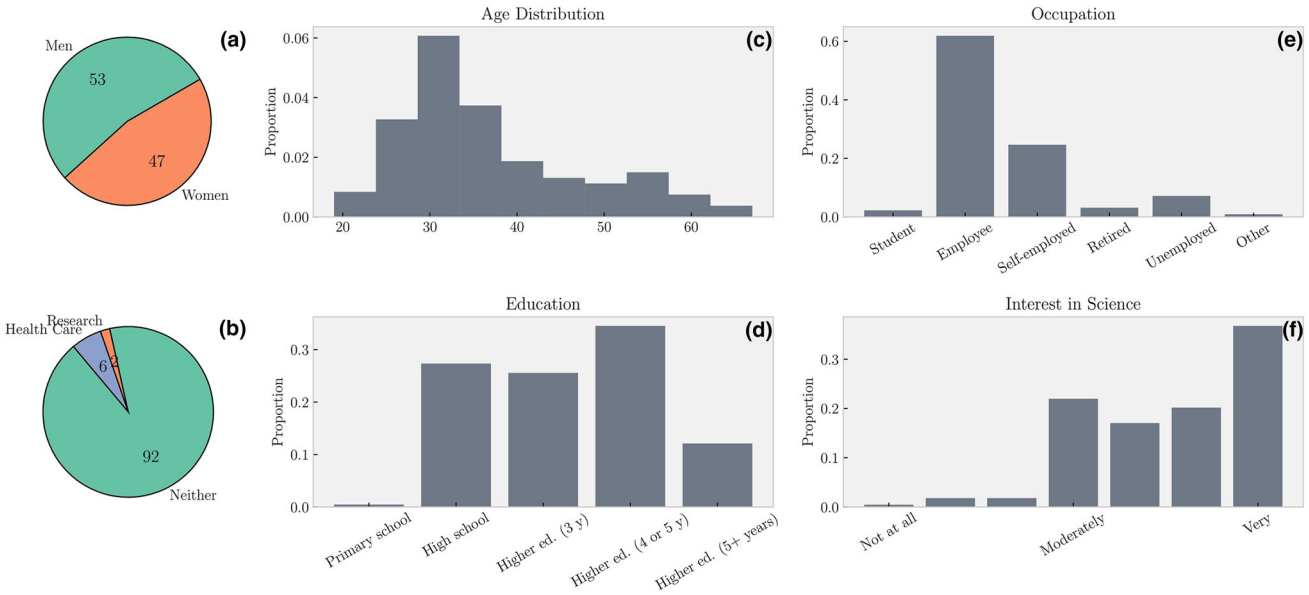

Fig. 1 Demographics of the respondents. (a) Gender, (b) Percentage of respondents working in health care or research, (c) Age distribution, (d) Distribution of educational level, (e) Occupation, (f) Interest level in science

did (Fig. 1b). A large proportion of respondents indicated they were employees or self-employed, with a least high-school education (See Fig. 1e,d. A more detailed breakdown can be found in the supporting material). Respondents considered themselves moderately to very interested in science (Fig. 1f).

4.2 Ranking Agreement

In the first part of the questionnaire, respondents were asked to rank two sets of actions according to the level they violate a patient's privacy or autonomy. We analyzed the agreement between respondents' rankings by calculating Kendall's W, both for actions violating privacy and actions violating autonomy. This statistic provides a measure of agreement between respondents ranging from 0 (no agreement) to 1 (complete agreement). We found Kendall's W coefficients of 0.37 and 0.64 for privacy and autonomy, respectively. Figure 2 depicts the agreement in ranking across correspondents.

4.3 Action Agreement

The second part of the questionnaire, respondents indicated which actions they deemed permissible in several scenarios leading a hypothetical patient to suffer from some conditions with different impacts. Figure 3a,b shows for each condition and each action the proportion of respondents deeming the action permissible. As there was considerable disagreement among respondents about the relative invasiveness of the actions, we also calculated these proportions as a function of the rank assigned to an action by each respondent (Fig. 3c,d).

Figure 3a–d reveals that for some combinations of actions and scenarios, there was a high level of agreement (proportions of participants close to 0 or 1, i.e., bright red or blue areas in Fig. 3a–d). However, for other combinations agreement was low (proportion of participants close to 0.5., i.e., dark areas in Fig. 3a–d).

To evaluate whether the respondents perceived the differences in the impact of the conditions, we ran a linear regression. This regression tested whether the probability an action was considered acceptable varied as a function of disease weight (Table 2. The disability weight was found to predict the acceptability of an action significantly. Also, the proportion of acceptable actions was higher for the actions about violations of privacy (see also Fig. 5 of supporting material) (Table 3).

5 Discussion

We asked 223 respondents to rank robotic actions according to their impact on the patient's autonomy and privacy. We found the agreement among respondents, as measured by Kendall's W was mediocre (privacy: W = 0.37; autonomy: W = 0.64, Fig. 2). When asking respondents to select actions they deemed permissible in 8 scenarios, differing in degree of the potential impact on the patient's well-being, the agreement was again mediocre (Fig. 3). The agreement did not increase after correcting for individual differences in the ranking of the actions (compare Fig. 3a,b and c,d). Hence, interpersonal disagreement about the relative impact of actions in itself did not explain the lack of agreement.

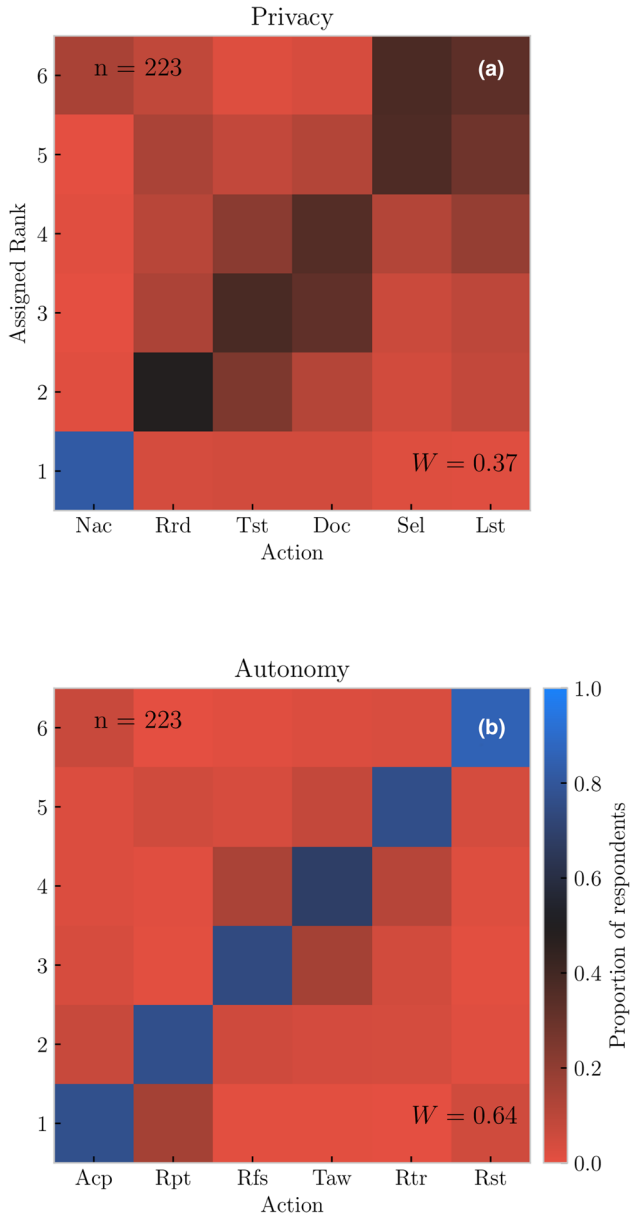

Fig. 2 Ranking agreement. Visualization of the contingency tables resulting from ranking each of the six actions violating privacy (a) and the six actions violating autonomy (b). In these matrices, both high values (close to 1, red) and low values (close to 0, blue) indicate high agreement among respondents. The respective values of Kendall's W (calculated across all values in the table) are denoted in the graphs

Despite the limited agreement among respondents, our data confirm that the empirical survey-based approach can serve as an efficient explorative tool. Indeed, although we found substantial disagreement for some actions, participants did agree on specific actions in particular contexts (the bright areas in Fig. 3a–d). For about 50% of action-disease combinations agreement was higher than 75%. Therefore, taking the role of CR developers, we argue the data can be translated into a number of boundaries for autonomous robot decisions. In particular, we list five behavioral rules for our hypothetical CR that can be extracted from the survey:

1. Repeating a request (Rpt) is considered very acceptable. Participants did not think this to violate a patient's autonomy (even though some authors have suggested it does, Deng [26]; Pontier and Hoorn [52]). Therefore, the robot should always repeat the question to the take the medication.
2. For all medical conditions, participants agreed that restraining a patient (Rst) is unacceptable. Therefore, the robot should never restrain a person.
3. Overall, taking no action (Nac, Acp) is less acceptable than the least invasive action (Rdf, Rpt). In particular, in the case of a patient who has acute schizophrenia, participants agreed that doing nothing (Nac) was unacceptable. Therefore, the robot should always take some action in this case (see also next item).
4. For the three most severe medical conditions, people agreed that some violation of privacy (Rdr, Tst and Doc) was acceptable. There was less agreement on these actions for conditions with lesser impact. Therefore, for a patient with a severe medical condition, the robot should record the decision and inform the doctor and/or a trusted person.
5. People seemed to agree that most violations of autonomy (Taw, Rtr, Rst) are unacceptable for the four least severe medical conditions. Less agreement was found for acute schizophrenia, severe depression, and severe Parkinson's disease. Therefore, a robot should never constrain the autonomy for a person with a less severe medical condition.

In addition to areas of agreement, it is interesting to note areas of disagreement between people. In particular, participants did not achieve a consensus about acceptable low-level privacy violations for less severe medical conditions. Nor did participants agree on the acceptability of the most invasive privacy violations for the most severe medical conditions (see dark regions Fig. 3a). People also did not agree on what violaions of autonomy are acceptable for cases pertaining to the most severe medical conditions. The areas of disagreement might require further finegrained inquiry to identify actions on which people agree (see also below).

These results show that the empirical approach can help in identifying agreed-upon (un)acceptable robot actions. Given the limitations of the top-down and bottom-up approaches discussed in the introduction and background section, we conclude that the empirical approach is a promising complementary avenue. Especially so since it a very rapid and cost-effective method to probe people's intuitions about ethical issues.

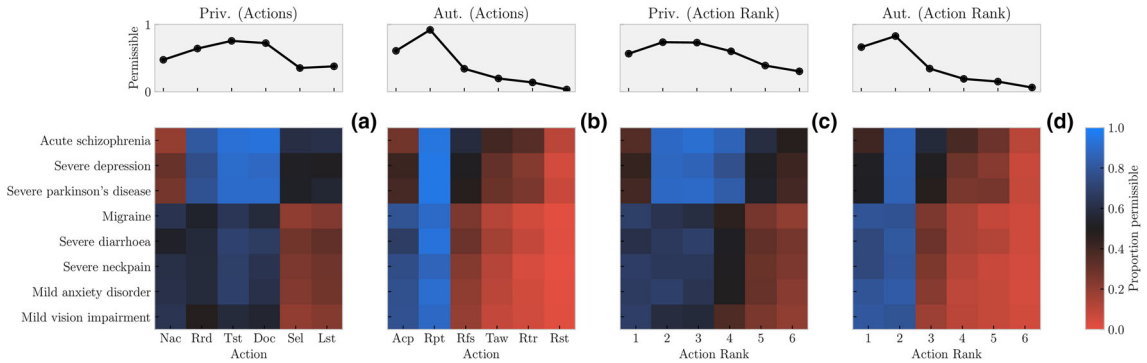

Fig. 3 Permissibility and agreement on actions. **Top panels** the average permissibility of each action or action rank across scenarios (i.e., average of panels a–d across rows). **a–d** Proportion for each of the privacy (a) and autonomy actions (b) listed in Table 1 for each medical condition. Panel c & d: similar, but for the rank each individual respondent assigned to each action in the first part of the survey. See Table 1 for the actions labels used in panels a–d

Table 3 Results of linear regression with proportion of permissible actions as dependent variable, and disability weights (Table 2) and domain (factor: privacy, autonomy) as independents

	Coef	SE	t	$P > \|t\|$	[0.025	0.975]
Intercept	0.3243	0.022	14.634	0.000	0.276	0.373
Domain	0.1270	0.031	4.053	0.002	0.059	0.195
Disability Weight	0.1322	0.048	2.731	0.018	0.027	0.238
Disability weight:domain	0.1460	0.068	2.133	0.054	-0.003	0.295

See Fig. 5 of supporting material for a visualization

Our study design might partly explain the limited agreement among respondents, and we suggest a potential route to maximize the informativity of our survey-based approach. In constructing the materials for our survey, we attempted to select a realistic scenario (treatment refusal) and implementable robot actions (Table 1). In doing so, we aimed at avoiding querying respondents on robotic behavior that pertains to highly unlikely scenarios and actions that are technically impractical (See [12,14,35], for examples and discussions). Nevertheless, our hypothetical situations leave many details open to the assumptions of the respondents. The limited agreement among respondents in certain areas might reflect differences in assumptions they made about the presented scenario, the robot, and its actions. Data from surveys querying the acceptance of CRs support this surmise.

In a survey conducted in 27 European countries, over 50% of the respondents indicated they wanted to see robots banned from providing care [11]. Also, almost 90% of respondents expressed being uncomfortable with the thought of robots caring for either children or the elderly. Nomura et al. [49] report high levels (24–42% of respondents) of anxiety associated with robots working in care and education roles. In contrast, studies assessing the acceptance of deployed CR systems have generally found positive attitudes towards robots (e.g., [41,45], and references therein). Moreover, data suggest that acceptance of CRs is multifaceted [13] and depends on the characteristics of the robot [51]. These results indicate that asking people whether they would accept a hypothetical robot might lead them to make (potentially, unrealistic) assumptions about the robots' capabilities and roles. In turn, this might lead to higher levels of skepticism. On the other hand, when faced with an actual CR fears and uncertaintity seem to dissapear and users are generally positive about their potential.

We expect respondents' agreement on the acceptability of actions to be higher for a specific, actual robot system operating in a particular setting. In other words, the agreement rates reported here might be limited by asking respondents to decide on the possible actions for a hypothetical robot operating in an underspecified situation. If this assumption were correct, this implies that the empirical approach should result in more clear-cut results and rules when evaluating real robots in concrete circumstances. In turn, this suggests that decision-makers and robot developers could use the empirical approach as an efficient way to explore acceptable boundaries for a robot's behavior once its behavioral repertoire is fixed and its operational context established.

The popular misconception that ethical behavior for machines only pertains to life and death situations plagues the emerging field of ethical robots. However, moral norms guiding practitioners are part of daily routine. For example, ethical norms regulate when and how medical staff share information or how they approach patients' failure to follow medical advice. Likewise, the behavioral routines of robots in

Table 4 Summary of the advantages of the empirical approach to selecting ethical rules for robots

Efficiency Opinions can be quickly and cheaply sourced by querying a large sample of people using surveys. Indeed, the cost per response is typically less than $1, and data collection usually takes less than 24h. Hence, internet-based surveys would give both academic and industrial robot developers a powerful tool to collect answers about acceptable behavior. Therefore, our approach speeds up the user testing phase of the R&D cycle
Maintaining human control The bottom-up approach, based on machine learning, might reflect biases inherent in the training data. In contrast, surveys can be designed to minimize bias and prejudice in the responses. Also, as rules are based on human responses, humans are in control of their design
Implementable rules The empirical approach can be used to evaluate specific behavioral options available to a CR, that are matched to the setting and role of the robot. Therefore, the empirical approach leads to directly implementable behavioral rules
Legitimacy The empirical approach mimics established social and democratic consensus processes used to identify, agree on, and collectively endorse policies. By approximating this concusses process, the empirical approach results in higher legitimacy of the ethical standards for robots
Transparency Using machine learning, the inferred rules of behavior might be opaque, even to the designer of the algorithm. Surveying opinions and deducing behavioral rules from the collected responses maintains transparency. Rules are made explicit before programming them into the robot. Rules can be communicated and (potentially, formally) verified. In contrast, rules extracted through machine learning are often opaque

care settings will include a multitude of implied minor ethical decisions. Robot developers will have to decide how privacy, autonomy, and well-being are weighted, ideally taking into account situational variables. Ultimately, this will determine whether the robots' behavior is acceptable to patients, family, and health care providers.

As outlined in the introduction, the field is lacking a validated method for establishing what behavior is deemed acceptable. The ability of robots to support, inform, and entertain patients continuously increases. Despite this, developers lack a systematic approach to deciding what patients, family caregivers, and healthcare providers deem acceptable.

Developing a robust design method for selecting rules and principles for CRs is essential for their success. As discussed by Alaiad and Zhou [2], an estimated 40% of IT innovations in healthcare have been abandoned, mostly due to a lack of understanding of the factors that lead to the acceptance of new technology—ensuring that CRs act ethically should increase the likelihood of patients, caregivers, and health professionals accepting them [13]. Studies have confirmed that a lack of trust and concerns about the ethical behavior of robots currently hamper the acceptance of CRs as carers [2,34]. Methods for selecting (and justifying) principles and rules to regulate robotic behavior might increase the success rate of innovative robot platforms and thereby accelerate development and progress in this area [26]. Here, we suggest and evaluate a promising design method for selecting rules and principles for CRs. We propose that the empirical approach can be an effective method that leads to directly implementable rules for CRs while maintaining human control and transparency (See tab. 4). Our approach might be relevant to other areas in which autonomous agents should behave ethically, such as self-driving cars Deng [26], and consider this as a pertinent direction for future research.

6 Conclusion

The limitations of current approaches to rule selection for ethical CRs warrant investigating other methods. We proposed a complementary survey-based method based on the input of multiple stakeholders. We argued that such an approach has several advantages, including the ability to assess practically relevant behavioral rules. For this to work, however, stakeholders should be able to come to a consensus about what is permissible. To explore the feasibility of our method, we surveyed people on some realistic robotic actions in a practically relevant scenario. From the data, we were able to derive five behavioral rules. Therefore, we conclude that surveys are a feasible, cost-effective, complimentary method to obtain transparent rules for CRs.

Acknowledgements Open access funding provided by Vienna University of Economics and Business (WU).

References

1. Abel D, MacGlashan J, Littman ML (2016) Reinforcement learning as a framework for ethical decision making. In: Workshops at the thirtieth AAAI conference on artificial intelligence

2. Alaiad A, Zhou L (2014) The determinants of home healthcare robots adoption: an empirical investigation. Int J Med Inform 83(11):825–840. https://doi.org/10.1016/j.ijmedinf.2014.07.003
3. Allen C, Smit I, Wallach W (2005) Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics Inf Technol 7(3):149–155. https://doi.org/10.1007/s10676-006-0004-4
4. Allen C, Wallach W, Smit I (2006) Why machine ethics? IEEE Intell Syst 21(4):12–17. https://doi.org/10.1109/MIS.2006.83
5. Anderson M, Anderson S, Armen C (2005) Towards machine ethics: implementing two action-based ethical theories. In: Fall symposium on machine ethics, pp 1–7
6. Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28(4):15–26. https://doi.org/10.1609/aimag.v28i4.2065
7. Anderson M, Anderson SL (2010) Robot be good. Sci Am 303(4):72–77
8. Arkin RC, Ulam P, Wagner AR (2012) Moral decision making in autonomous systems: enforcement, moral emotions, dignity, trust, and deception. Proc IEEE 100(3):571–589
9. Beauchamp TL, Childress JF (2001) Principles of biomedical ethics. Oxford University Press, New York
10. Bemelmans R, Gelderblom GJ, Jonker P, de Witte L (2012) Socially assistive robots in elderly care: a systematic review into effects and effectiveness. J Am Med Dir Assoc 13(2):114–120. https://doi.org/10.1016/j.jamda.2010.10.002
11. Bogue R (2014) The future of robotics in Europe. Industrial Robot: An International Journal 41(6):487–492. https://doi.org/10.1227/NEU.0b013e318271ff20
12. Bonnefon J-F, Shariff A, Iyad R (2015) The social dilemma of autonomous vehicles. Science 1080(2013):1573–1576. https://doi.org/10.1126/science.aaf2654
13. Broadbent E, Stafford R, MacDonald B (2009) Acceptance of healthcare robots for the older population: review and future directions. Int J Soc Robot 1(4):319–330. https://doi.org/10.1007/s12369-009-0030-6
14. Brooks R (2017) Unexpected consequences of self driving cars. https://rodneybrooks.com/unexpected-consequences-of-self-driving-cars/
15. Brundage M (2014) Limitations and risks of machine ethics. J Exp Theor Artif Intell 26(3):355–372. https://doi.org/10.1080/0952813X.2014.895108
16. Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. https://doi.org/10.1177/2053951715622512
17. Burtness JH (1999) Consequences: morality, ethics, and the future. Fortress Press, Minneapolis
18. Callahan D (2012) The roots of bioethics: health, progress, technology, death. Oxford University Press, Oxford
19. Cassidy B, Blessing JD (2007) Ethics and professionalism: a guide for the physician assistant. FA Davis, Duxbury
20. Castro J (2016) A bottom-up approach to machine ethics. In: Proceedings of the artificial life conference 2016. MIT Press, Cancun, Mexico, pp 712–719. ISBN 978-0-262-33936-0. https://doi.org/10.7551/978-0-262-33936-0-ch113
21. Chadwick RF, Schroeder D (2002) Applied ethics: critical concepts in philosophy, vol 6. Taylor & Francis, Abingdon
22. Char DS, Shah NH, Magnus D (2018) Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 378(11):981–983. https://doi.org/10.1056/NEJMp1714229
23. Checkel JT (1999) Social construction and integration. J Eur Public Policy 6(4):545–560. https://doi.org/10.1080/135017699343469
24. Copp D (2005) The Oxford handbook of ethical theory. Oxford University Press, Oxford
25. Côté S, Piff PK, Willer R (2013) For whom do the ends justify the means? Social class and utilitarian moral judgment. J Personal Soc Psychol 104(3):490–503. https://doi.org/10.1037/a0030931 ISSN 00223514
26. Deng B (2015) Machine ethics: the robot's dilemma. Nature 523(7558):20
27. Ding J, Lim Y-J, Solano M, Shadle K, Park C, Lin C, Hu J (2014) Giving patients a lift-the robotic nursing assistant (rona). In: 2014 IEEE international conference on technologies for practical robot applications (TePRA). IEEE, pp 1–5
28. Everett Jim AC, Pizarro David A, Crockett MJ (2016) Inference of trust worthiness from intuitive moral judgments. J Exp Psychol Gen 145(6):772–787. https://doi.org/10.1037/xge0000165 ISSN 00963445
29. Feil-Seifer D, Mataric MJ (2005) Defining socially assistive robotics. In: 9th international conference on rehabilitation robotics, 2005. ICORR 2005. IEEE, pp 465–468
30. Feil-Seifer D, Matarić MJ (2007) Socially assistive robotics. Robot Autom Mag IEEE 18(1):24–31. https://doi.org/10.1109/MRA.2010.940150
31. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42(3):143–166
32. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G (2018) Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 178(11):1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763 ISSN 2168-6106
33. Gips J (1995) Towards the ethical robot. Android Epistemol, pp 243–252
34. Glende S, Conrad I, Krezdorn L, Klemcke S, Krätzel C (2015) Increasing the acceptance of assistive robots for older people through marketing strategies based on stakeholder needs. Int J Soc Robot 8(3):355–369. https://doi.org/10.1007/s12369-015-0328-5
35. Goodall NJ (2016) Away from trolley problems and toward risk management. Appl Artif Intell 30(8):810–821. https://doi.org/10.1080/08839514.2016.1229922 ISSN 10876545
36. Greene JD (2016) Ethics. Our driverless dilemma. Science (New York, NY) 352(6293):1514–1515. https://doi.org/10.1126/science.aaf9534
37. Guarini M (2006) Particularism and the classification and reclassification of moral cases. IEEE Intell Syst 21(4):22–28. https://doi.org/10.1109/MIS.2006.76
38. Honarvar AR, Ghasem-Aghaee N (2009) An artificial neural network approach for creating an ethical artificial agent. In: 2009 IEEE international symposium on computational intelligence in robotics and automation—(CIRA), Daejeon, Korea (South). IEEE, pp 290–295. https://doi.org/10.1109/CIRA.2009.5423190
39. Islam SR, Kwak D, Kabir MH, Hossain M, Kwak K-S (2015) The internet of things for health care: a comprehensive survey. IEEE Access 3:678–708
40. Kanemura A, Morales Y, Kawanabe M, Morioka H, Kallakuri N, Ikeda T, Miyashita T, Hagita N, Ishii S (2013) A waypoint-based framework in brain-controlled smart home environments: brain interfaces, domotics, and robotics integration. In: 2013 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 865–870
41. Koceski S, Koceska N (2016) Evaluation of an assistive telepresence robot for elderly healthcare. J Med Syst 40(5):1–8. https://doi.org/10.1007/s10916-016-0481-x
42. Kulkarni A, Sathe S (2014) Healthcare applications of the internet of things: a review. Int J Comput Sci Inf Technol 5(5):6229–6232
43. Levine C (2012) My father won't take his meds. AARP.org. https://www.aarp.org/home-family/caregiving/info-08-2012/father-wont-take-his-medication.html
44. Mackworth AK (2011) Architectures and ethics for robots. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 204–221
45. Mast M, Burmester M, Kruger K, Fatikow S, Arbeiter G, Graf B, Kronreif G, Pigini L, Facal D, Qiu R (2012) User-centered design of

a dynamic-autonomy remote interaction concept for manipulation-capable robots to assist elderly people in the home. J Hum Robot Interact 28:96–118. https://doi.org/10.5898/JHRI.1.1.Mast
46. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc. https://doi.org/10.1177/2053951716679679
47. Mukai T, Hirano S, Yoshida M, Nakashima H, Guo S, Hayakawa Y (2011) Whole-body contact manipulation using tactile information for the nursing-care assistant robot riba. In: 2011 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, p 2445–2451
48. Murphy RR, Woods DD (2009) Beyond Asimov: the three laws of responsible robotics. IEEE Intell Syst. https://doi.org/10.1109/MIS.2009.69
49. Nomura T, Kanda T, Suzuki T, Kato K (2009) Age differences and images of robots: social survey in Japan. Interact Stud 10(3):374–391. https://doi.org/10.1075/is.10.3.05nom
50. Picard RW, Picard R (1997) Affective computing, vol 252. MIT Press, Cambridge
51. Pino M, Boulay M, Jouen F, Rigaud A-S (2015) Are we ready for robots that care for us? Attitudes and opinions of older adults toward socially assistive robots. Front Aging Neurosci 7:141. https://doi.org/10.3389/fnagi.2015.00141
52. Pontier M, Hoorn J (2012) Toward machines that behave ethically better than humans do. In: Proceedings of the annual meeting of the cognitive science society, vol 34
53. Russell S, Daly J, Hughes E, Hoog CO (2003) Nurses and 'difficult' patients: negotiating non-compliance. J Adv Nurs 43(3):281–287
54. Salomon JA, Haagsma JA, Davis A, de Noordhout CM, Polinder S, Havelaar AH, Cassini A, Devleesschauwer B, Kretzschmar M, Speybroeck N, Murray CJL, Vos T (2015) Disability weights for the global burden of disease 2013 study. Lancet Global Health 3(11):e712–e723. https://doi.org/10.1016/S2214-109X(15)00069-8
55. Scheutz M (2017) The case for explicit ethical agents. AI Mag 38(4):57–64. https://doi.org/10.1609/aimag.v38i4.2746
56. Sharkey N (2008) The ethical frontiers of robotics. Science 322(5909):1800–1801
57. Sharkey NE, Sharkey AJC (2011) The rights and wrongs of robot care. In: Lin P, Bekey G, Abney K (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp 267–282
58. Sparc. Robots that may help you in your silver age, 2016. http://robohub.org/robots-that-may-help-you-in-your-silver-age/
59. Ueda J, Kurita Y (2016) Human modeling for bio-inspired robotics: mechanical engineering in assistive technologies. Academic Press, New York
60. van Rysewyk SP, Pontier M (2015) A hybrid bottom-up and top-down approach to machine medical ethics: theory and data. In: van Rysewyk SP, Pontier M (eds) Machine medical ethics, vol 74. Springer, Cham, pp 93–110
61. Vanderelst D, Winfield A (2018) An architecture for ethical robots inspired by the simulation theory of cognition. Cogn Syst Res 48:56–66
62. Wallach W, Allen C (2009) Top-down morality. In: Moral machines: teaching robots right from wrong. Oxford University Press, New York. https://doi.org/10.1093/acprof:oso/9780195374049.003.0007
63. Winfield AFT (2014) Robots with internal models: a route to self-aware and hence safer robots. In: Pitt J (ed) The computer after me: awareness and self-awareness in autonomic systems, 1st edn. Imperial College Press, London. ISBN 9781783264179
64. Winfield AFT, Blum C, Liu W (2014) Towards an ethical robot: internal models, consequences and ethical action selection. In: Adv Auton Robot Syst, pp 85–96. Springer, Cham

Dieter Vanderelst obtained an MSc in Theoretical Psychology (Ghent University, Belgium) and an MSc in Artificial Intelligence (Leuven University, Belgium). He got his Ph.D. in Biology (2012, University of Antwerp, Belgium). As a postdoc, he worked at the University of Bristol (UK) as a Marie Curie Fellow and the Bristol Robotics Laboratory (UK). He joined the University of Cincinnati in Augst 2016 as an assistant professor with a joint appointment in the Psychology, Biological Sciences, Electrical Engineering & Computing Systems, and Mechanical & Materials Engineering Departments.

Jurgen Willems is Professor for Public Management and Governance at the Vienna University of Economics and Business (WU Wien). His teaching and research cover a variety of topics on citizen-state and citizen-society interactions. Concrete areas of interest are: Changing civic engagement and its new challenges for public policy and management; Network governance for within-sector and cross-sector collaborative ecosystems; and the impact of new technologies on citizen-state and citizen-society interactions.