Implementing Pro-social Rule Bending in an Elder-care Robot Environment

Rajitha Ramanayake[0000−0001−9903−0493] and Vivek Nallur[0000−0003−0447−4150]

School of Computer Science, University College Dublin, Belfield, Republic of Ireland
rajitha.ramanayakemahantha@ucdconnect.ie
vivek.nallur@ucd.ie

Abstract. Many ethical issues arise when robots are introduced into elder-care settings. When ethically charged situations occur, robots ought to be able to handle them appropriately. Some experimental approaches use (top-down) moral generalist approaches, like Deontology and Utilitarianism, to implement ethical decision-making. Others have advocated the use of bottom-up approaches, such as learning algorithms, to learn ethical patterns from human behaviour. Both approaches have their shortcomings when it comes to real-world implementations. Human beings have been observed to use a hybrid form of ethical reasoning called Pro-Social Rule Bending, where top-down rules and constraints broadly apply, but in particular situations, certain rules are temporarily bent. This paper reports on implementing such a hybrid ethical reasoning approach in elder-care robots. We show through simulation studies that it leads to better upholding of human values such as autonomy, whilst not sacrificing beneficence.

Keywords: Elder-care robots · Machine ethics · Ethical decision making · Ethical governor · Rule bending

1 Introduction

The world has a growing aged population. Many have proposed the use of robots, as a solution to the rising problem of caring for the elderly. As a result, many elder-care robots with different abilities are available in the market [9]. Empirical studies have concluded that the stakeholders in the elder-care environment find many ethical concerns regarding the delegation of work from human care-workers to robots [9]. Hence, it is a common view that these robots should have the capacity to act ethically, in ethically charged situations in their work environment. Ramanayake and Nallur [6] argued that some of these ethical concerns, such as concerns regarding privacy, wellbeing, autonomy, and availability, can be solved by better technical implementations that take concerns of various stakeholders (e.g., patients, careworkers, family, etc) into consideration. Any decision-making mechanism used by the robot, apart from being functionally adequate, must evaluate the impact of the decision on the ethical concerns as well.

The field of machine-implemented ethics can roughly be categorised into three, based on the engineering approach of the ethical decision making system.

These three approaches are namely: Top-down, Bottom-up, and Hybrid [13]. Many traditional generalist ethical theories of the world (e.g., deontological ethics, legal codes, and utilitarian ethics) and the computational systems that adapted those follow top-down decision making. In this approach, the designers of the systems try to foresee decision points, and decide what is ethical (or not) (e.g. [3]), and programme them into the system. Most current implementations of this appraoch use logic frameworks or simulations to reason about the ethical acceptability of a particular behaviour. In a bottom-up approach, the system is designed with social and cognitive processes which interact with each other, and the environment. Using these interactions, or from supervision, it is expected to learn what is ethical (or not) and behave accordingly. Hence, ethical decisions made by these systems are not guided by any ethical theory. Implementations that follow this approach use algorithms such as social choice theory and voting based methods, and Artificial Neural Networks to capture ethical patterns of the environments [10].

The main shortcoming of the top-down approach is that it can only guarantee ethical behaviour in relatively small and closed systems where the designers can know all the possible states of the system. In contrast, systems designed through the bottom-up approach require complex cognitive and social process models, a large amount of reliable and accurate data, and a comprehensive knowledge model of the world to learn intricate social constructs such as ethics [5, 8]. The hybrid approach to implementing ethical machines is considered to be a good alternative to overcome the shortcomings of the other approaches [13]. The key idea behind this approach is to combine the flexibility and evolving nature of the bottom-up approach with the value, duty and principle-oriented nature of the top-down approach to create a better, more reliable system.

The domain of care does not admit neat ethical theorisation. Kantian and rights-based ethics, and utilitarian ethics have been pointed out as being inadequate [5] in real-world care settings. However, most computational ethics implementations in robots in literature [3, 11, 12] use such theorisation. Hence, some argue that a good ethical reasoner should be able to step *out* of existing ethical theoretical frameworks, but *only* when necessary [1]. Pro-social rule bending (PSRB) has been identified by Morrison [4] as the mechanism by which human beings (in other contexts) step outside of rigid ethical constraints. Therefore, it has been suggested that PSRB could be a good (and unexplored) contender for real-world ethical dilemmas [8] (such as scenarios introduced in [7]).

This paper reports on an implementation of PSRB and how it affects decision-making in a specific dilemma, that affects the elderly in an assisted living environment. The presented ethical governor model uses expert knowledge and case-based reasoning (CBR) to analyse rule-bending behaviours, and contest the top-down rule system's decisions when required. By doing so it makes the rule system behave more desirably when it encounters infrequent circumstances [6]. Our model of PSRB capable ethical governor employs the hybrid approach in the sense that we use the knowledge acquired bottom-up to contest the top-down rules that are programmed into the system at design time.

2 An Implementation of PSRB Capable Ethical Governor

As discussed in the previous section, this paper attempts to bring together, two novel concepts: concern for human autonomy as well as implement a hybrid mechanism to perform ethical reasoning. As Ramanayake and Nallur point out [7], there are several small inter-personal scenarios in daily life, which present ethically challenging decision points. Out of these, we pick a dilemma that shows the conflict between autonomy and human well-being, called the *Bathroom Dilemma*. We simulate an elder-care robot caught in this dilemma, and the particular way in which a PSRB-capable ethical governor picks an action. We contrast it with the same robot, using Deontological as well as Utilitarian reasoning mechanisms.

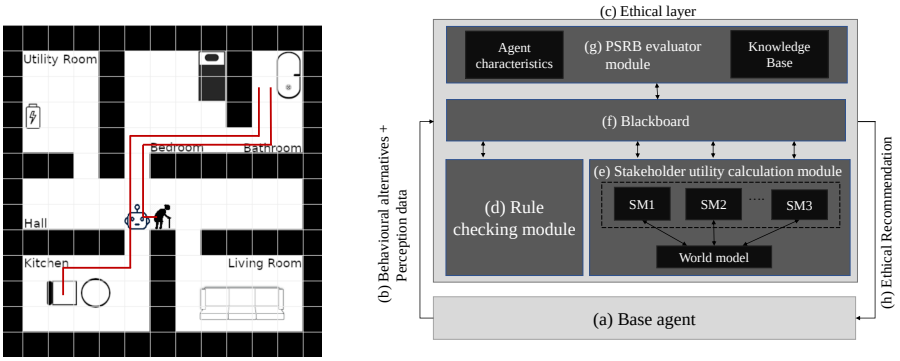
Bathroom Dilemma An elder-care robot is assigned to an elderly resident, who lives alone. The main task of the robot is to follow the resident around the house, and record activities of daily living. These recordings will be used to identify any cognitive issues of the resident. The robot has the ability to identify emergencies involving the resident. Also, when its battery power is low, it can autonomously go to the charging station. The robot is connected to a database that contains the resident's history and current health status.

In this dilemma, the resident goes into the bathroom. However, before going in, the resident commands the robot not to follow them into the bathroom. The average time the resident stays in the bathroom is 10 minutes with a 5-minute standard deviation. In this instance, the resident stays in the bathroom for over 15 minutes. This robot has only three actions to choose from. 1) Stay outside the bathroom 2) Go inside the bathroom or 3) Go to the charging station. If the robot stays outside, the resident's wellbeing is at risk. However, going inside will undermine the resident's autonomy. Other variables such as the *time since the resident entered the bathroom*, the resident's *health*, the resident's *medical history* and the *battery level* of the robot can affect the robot's decision.

2.1 The Simulation Environment

We created a virtual simulation environment of an ambient assisted living (AAL) space using modified *MESA* agent-based modelling framework [2]. The simulation environment is a 13×13 grid which contains a resident and the robot. The robot agent can only see objects in a 3-step radius and cannot see through walls. While in the charging state, the robot will charge 3 units of power per step and in every other state it will spend 0.2 units. The environment allows resident agents to move anywhere in the grid other than the locations of the walls and the robot.

The Human Agent We define the human agent in the environment as a path-following agent and it can give instructions to the robot agent. Both instructions and the path can be given as user inputs to the simulator. However, when the



(a) The simulation environment and path of the human agent

(b) Architecture of the rule-bending agent

Fig. 1: The Simulation Environment and the PSRB-capable Governor Architecture

robot is blocking the human's path, the human agent will give the `move_away` instruction to the robot autonomously.

2.2 Architecture for a Pro-social Rule Bending Agent

We use the PSRB-capable computational architecture introduced in [6] illustrated in Figure 1b. This is the first implementation of a PSRB-capable agent that we are aware of. In this section, we will briefly explain the architecture and its main elements (shown as (a), (b), ... in Figure 1b).

The Monitoring Robot Agent The base agent (a) is an autonomous agent that collects perception data from the environment and decides its next move, at every step. Its main goal is to follow the human agent assigned to it. The robot agent also can go to the charging station autonomously. It has the ability to follow instructions,

1. `move_away` - Triggers behaviour of moving away.
2. `do_not_follow_to__<room_name>` - Restrict moving to the `<room_name>`
3. `continue` (following) - Remove any restrictions posed by instruction 2

We call these instructions *Instruction 1, 2, 3* from here onward. The robot only accepts these commands when the command giver can be seen. The robot agent has several behaviour priorities. The highest priority is going to the charge station when the battery is less than 5%. Its next priority is to follow the resident. When it does not see the resident it tries to minimise the distance between itself and the last seen location of the resident. The robot agent generates a number of behavioural alternatives in a given situation, and passes them and

the perception data (b) to the ethical layer. The ethical layer recommends one or more behavioural alternatives to the base agent (h). When there is more than one recommendation, the base agent chooses one among them considering its behaviour priorities.

Rule Checking Module Rule checking module (d) checks the permissibility of each action according to the rule set and stores the results with the `IDs` of the rules broken in the blackboard. This implementation follows three rules.

1. When *instruction 2* is received, it is not permissible to go to the `<room_name>` until the same resident issues *instruction 3*.
2. When *instruction 1* is received, it is not permissible to not move away.
3. It is not permissible to go to the charge station when the battery percentage is more than 25%.

Stakeholder Utility Calculation Module This module (e) calculates the utilities for the values 'Autonomy', 'Wellbeing' and 'Availability', for each stakeholder in every step and for every behaviour alternative. It stores these calculated values on the blackboard. *Note*: The functions used in this implementation are created for the simulator. A real world implementation will need more accurate utility functions.

Autonomy We define the autonomy utility for this scenario considering two factors: whether the robot obeys the resident's instructions, and whether it physically limits the user from doing something. For this implementation, we consider the latter to be the biggest violation.

$$Au_i = \begin{cases} -1: & \text{if the resident is physically restrained by the robot} \\ -0.7: & \text{if the robot disobeys a resident instruction} \\ 0: & \text{if no instructions given} \\ 1: & \text{if the robot obeys a resident instruction} \end{cases} \tag{1}$$

Wellbeing We consider that the resident's wellbeing is at its highest as long as the robot sees the resident and the resident is not in danger. The longer the robot loses sight of the resident (T), the lower the wellbeing score, because of the uncertainty of the patient's state. Other variables such as the average and standard deviation time in the room r: $\bar{T}_r$ and σ_r, the number of emergencies in the past: p, and the resident health score: $h \in [0, 1]$ controls the gradient of the utility function. However, when the behaviour is `go to last seen`, the wellbeing utility represents the wellbeing-centred nature of the action and gives higher utility to encourage such behaviour. Wellbeing utility of behaviour i is W_i,

$$t = \frac{1}{(1+e^{p-2})} \quad (2) \qquad W_i = \begin{cases} 0.7: & \text{if } i = \text{going to last seen location} \\ \frac{2}{\left(1+e^{\frac{h(1-t)(T-(\bar{T}_r+\sigma_r))}{2}}\right)} - 1: & \text{else} \end{cases} \tag{3}$$

Availability This utility declines with the robot's battery level b. However, in situations where behaviour $i =$ `go to the charge station` and the battery is low, the utility gives a positive boost to represent the 'availability maximising' nature of that behaviour. Availability utility is Av_i,

$$y = \frac{-28.125}{b+12.5} + 1.25 \quad (4) \qquad Av_i = \begin{cases} y + abs(y): & \text{if } i = \text{go to charge station} \\ & \textbf{AND } y < 0.4 \\ y: & else \end{cases} \tag{5}$$

PSRB Evaluator Module The PSRB evaluator module has two main components: *Agent Character* and *Knowledge Base*

Agent Character There are many character traits that affects PSRB behaviour (i.e., risk propensity, robot's autonomy, etc.) [6]. The person/organisation authorised to set up the robot can define these character traits for the robot. For simplicity, we use value preferences as the only character variable. One can set a number between [1,10] for each value (i.e, autonomy (C_{au}), wellbeing (C_w) and availability (C_{av})) which will reflect the agent's precedence regarding said values. In this instance they are set to $C_w = 9$, $C_{au} = 3$ and $C_{av} = 3$, indicating that the robot's character is to prioritise wellbeing when needed.

Knowledge Base The task of the knowledge base is to return the absolute or approximate expert opinion, given a context. To this end, we use Case-Based Reasoning (CBR). Implicit explainability, traceability, and the ability to work with incomplete queries and data are the main reasons we chose a CBR system. The latter is crucial in these types of scenarios because some cases might have additional variables that others do not have (e.g., last-seen location, last-seen time). The system uses a mix of perception data, calculated utilities and the behaviour to represent a case. For each expert opinion, the intention is also recorded. When queried, the knowledge base returns the opinion on the acceptability of the behaviour and the intention behind it. An experienced elder-care practitioner was consulted to validate the knowledge base used. This implementation uses the K-Nearest Neighbours algorithm with $K = 3$ and inverse distance voting function when $distance > 0.2$ as the retrieval algorithm. When the $distance \leq 0.2$, it uses 5 as the weight of the instance.

3 Comparison of PSRB Behaviour with Other Approaches

3.1 Experiment Setup

Comparing Agents We implemented two agents based on two ethical frameworks that are commonly used in existing systems.

$Agent_D$ which pursues the deontological ethics approach adheres to the rules specified in section 2.2. These rules are not specifically designed to perfectly handle every situation in this environment. However, this is intentional and done to acknowledge the challenge of designing a comprehensive rule set that can account for all contingencies in a complex environment. This limitation is also noticeable in many real-world rule systems, such as road rules and healthcare rules [1].

The second, an act-utilitarian agent ($Agent_U$) uses the following formula to aggregate the utilities ($U_{follower}$), and calculate the desirability of the behaviour. For this implementation, we give Au_i and W_i utilities the same weight. However, the weight of Av_i altered from the other two to increase its effects on total utility when the value of Av_i is low. The desirability of a behaviour is set to the maximum when $U_{follower}$ is higher than 0.5 and the lowest when it is less than 0, to reduce noice. In every other case, we set the desirability score to $U_{follower}$ so that it can be compared with the other behaviours.

$$U_{follower} = \frac{Au_i + W_i + Av_i e^{-Av_i}}{3} \tag{6}$$

$$D_i = \begin{cases} 1: & U_{follower} > 0.5 \\ U_{follower}: & 0.5 \geq U_{follower} \leq 0 \\ 0: & U_{follower} < 0 \end{cases} \tag{7}$$

The source code of the three agent implementations and the experiments can be found in an online public repository[1].

3.2 Cases With and Without Dilemma

Base Conditions For the normal case, $b = 100$, $\bar{T}_{bathroom} = 20$, $\sigma_{bathroom} = 10$, $p = 0$, and $h = 1$. In the experiment environment, the time is measured in the number of steps. For these cases, we use the 1 *step* = 30 *seconds* conversion function.

Case 1 The resident takes the path shown in Figure 1a. First, they start moving from `(6,5)` grid location to the bathroom, along the indicated path. Before entering, they issue `do_not_follow_to__bathroom` from the grid location `(10, 8)` (just after the resident enters the bathroom). The resident remains in the bathroom for 20 steps (10 minutes)(=$\bar{T}_{bathroom}$). Afterwards, they return to the kitchen via the indicated path while issuing *instruction 2* when they reach the grid point `(8,8)`.

[1] https://bit.ly/3R4Xbid

Case 2 The same as Case 1, however, the resident does not come out of the bathroom.

Case 3 The same as Case 1, however, the resident takes longer than normal to exit the bathroom (40 steps (=20 minutes)).

Case 4 The same as Case 1, however, the robot's battery level is extremely low. *Availability* starts to conflict with *Well-being*.

Case 5 The same as Case 1, however, the resident has a history of injury inside the bathroom.

A summary of all the different cases and decisions made by different moral reasoning implementations can be seen in Table 1.

Table 1: Cases and Behaviour With Differing Moral Reasoning Mechanisms

Case ID	Time spent in the bathroom	Circumstance	Agent_D	Agent_U	Agent_PSRB
1	10 minutes (20 steps)	Normal	Staying out	Staying out	Staying out
2	∞	Normal	Staying out	Go in at *step 271*	Go in at *step 43*
3	20 minutes (40 steps)	Normal	Staying out	Staying out	Go in at *step 43*
4	10 minutes (20 steps)	Low Battery ($b = 8$)	Go to charge at *step 26*	Go to charge at *step 1* Go to last seen at *step 9* Go to charge at *step 36* Go to last seen at *step 44*	Go to charge at *step 26* Go to last seen at *step 47*
5	10 minutes (20 steps)	History of emergencies ($p = 3$)	Staying out	Go in at *step 82*	Go in at *step 23*

4 Discussion of Behaviour

Case 1 This demonstrates that for most daily living activities, that are carefully considered during design time, all three robots perform as expected.

Case 2 The $Agent_D$ illustrates the consequences of the lack of an implicit rule about the time duration that is acceptable for the resident to stay in. One could argue that the ethical governor limited the base agent's full potential by precluding the base agent's default behaviour. The $Agent_U$ managed to allow this default behaviour after a long period of waiting. Nevertheless, both of these agents might not be able to send a life-saving alert to a human care-worker or an ambulance on time. The $Agent_{PSRB}$, on the other hand, triggered a PSRB behaviour enabling the default behaviour, around the time $\sim (\bar{T}_{bathroom} + \sigma_{bathroom})$, which is more suitable in the given circumstance. This result demonstrates that this approach can add flexibility and enhance otherwise rigid governing systems, empowering the bottom-up knowledge collected through user feedback and observing expert behaviour.

Case 3 This case shows that PSRB is not infallible. $Agent_{PSRB}$ acts cautiously, compared to the other agents, and checks on the resident. By doing so it violates the resident's autonomy without any gain. In this case, $Agent_D$ and $Agent_U$ performed better than $Agent_{PSRB}$. The main reason for this is the partially observable environment chosen in the experiment. We believe that partially observable environments, in general, are more representative of the real-world.

Cases 4 and 5 showcase how well the PSRB capable system works compared to traditional systems when handling infrequent cases. In Case 4, the $Agent_U$ abandons the resident as soon as it needs recharging. $Agent_D$ again blocked the default behaviour to uphold the resident's autonomy, by refraining from checking on the resident. However, $Agent_{PSRB}$ manages to stay close to the resident as much as it can and then go to the charging station. PSRB evaluator refusing the knowledge base suggestions (to *go to charge station* from step 13), in this instance shows that $Agent_{PSRB}$ also regulates itself well in this scenario to not overdo PSRB. Once sufficiently charged, the PSRB system again activates and allows the robot to check on the resident by moving towards the resident's last seen location. In case 5, $Agent_{PSRB}$ identified the change in context and acted accordingly. $Agent_U$ also reduced the wait to go in and check on the resident. However, it is still not nearly close enough to the $\bar{T}_{bathroom}$.

The behaviour shown from cases 1-5 demonstrated the enhancements a PSRB behaviour brings to rule-based and utility-based ethics approaches. We do not claim that the PSRB is a new school of ethics. Rather, we consider it an enhancement to the existing approaches that allow them to change the decision-making criteria, for the sole purpose of increasing social welfare.

References

1. Bench-Capon, T., Modgil, S.: Norms and value based reasoning: justifying compliance and violation. Artificial Intelligence and Law **25**(1), 29–64 (2017). https://doi.org/10.1007/s10506-017-9194-9
2. Kazil, J., Masad, D., Crooks, A.: Utilizing Python for Agent-Based Modeling: The Mesa Framework. In: Thomson, R., Bisgin, H., Dancy, C., Hyder, A., Hussain, M.

(eds.) Social, Cultural, and Behavioral Modeling. pp. 308–317. Springer International Publishing, Cham (2020)
3. Kim, J.W., Choi, Y.L., Jeong, S.H., Han, J.: A Care Robot with Ethical Sensing System for Older Adults at Home. Sensors **22**(19), 7515 (Jan 2022). https://doi.org/10.3390/s22197515, https://www.mdpi.com/1424-8220/22/19/7515, number: 19 Publisher: Multidisciplinary Digital Publishing Institute
4. Morrison, E.W.: Doing the job well: An investigation of pro-social rule breaking. Journal of Management **32**(1), 5–28 (2006). https://doi.org/10.1177/0149206305277790
5. Pirni, A., Balistreri, M., Capasso, M., Umbrello, S., Merenda, F.: Robot Care Ethics Between Autonomy and Vulnerability: Coupling Principles and Practices in Autonomous Systems for Care. Frontiers in Robotics and AI **8**, 654298 (Jun 2021). https://doi.org/10.3389/frobt.2021.654298, https://www.frontiersin.org/articles/10.3389/frobt.2021.654298/full
6. Ramanayake, R., Nallur, V.: A Computational Architecture for a Pro-Social Rule Bending Agent. In: First International Workshop on Computational Machine Ethics held in conjunction with 18th International Conference on Principles of Knowledge Representation and Reasoning KR 2021 (CME2021) (Nov 2021). https://doi.org/10.5281/ZENODO.6470437, https://zenodo.org/record/6470437
7. Ramanayake, R., Nallur, V.: A Small Set of Ethical Challenges for Elder-Care Robots. In: Hakli, R., Mäkelä, P., Seibt, J. (eds.) Frontiers in Artificial Intelligence and Applications. IOS Press (Jan 2023). https://doi.org/10.3233/FAIA220605, https://ebooks.iospress.nl/doi/10.3233/FAIA220605
8. Ramanayake, R., Wicke, P., Nallur, V.: Immune moral models? Pro-social rule breaking as a moral enhancement approach for ethical AI. AI & SOCIETY (May 2022). https://doi.org/10.1007/s00146-022-01478-z, https://link.springer.com/10.1007/s00146-022-01478-z
9. Sharkey, A., Sharkey, N.: Granny and the robots: Ethical issues in robot care for the elderly. Ethics and Information Technology **14**(1), 27–40 (Jul 2012). https://doi.org/10.1007/s10676-010-9234-6, https://link.springer.com/article/10.1007/s10676-010-9234-6, publisher: Springer
10. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A.: Implementations in Machine Ethics: A Survey. ACM Computing Surveys **53**(6) (2021). https://doi.org/10.1145/3419633, https://doi.org/10.1145/3419633, arXiv: 2001.07573
11. Van Dang, C., Tran, T.T., Gil, K.J., Shin, Y.B., Choi, J.W., Park, G.S., Kim, J.W.: Application of soar cognitive agent based on utilitarian ethics theory for home service robots. In: 2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI). pp. 155–158 (Jun 2017). https://doi.org/10.1109/URAI.2017.7992698
12. Vanderelst, D., Winfield, A.: An architecture for ethical robots inspired by the simulation theory of cognition. Cognitive Systems Research **48**, 56–66 (2018). https://doi.org/10.1016/j.cogsys.2017.04.002, https://doi.org/10.1016/j.cogsys.2017.04.002, publisher: The Authors
13. Wallach, W., Allen, C., Smit, I.: Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. AI and Society **22**(4), 565–582 (Apr 2008). https://doi.org/10.1007/s00146-007-0099-0