A Virtuous Care-Robot?

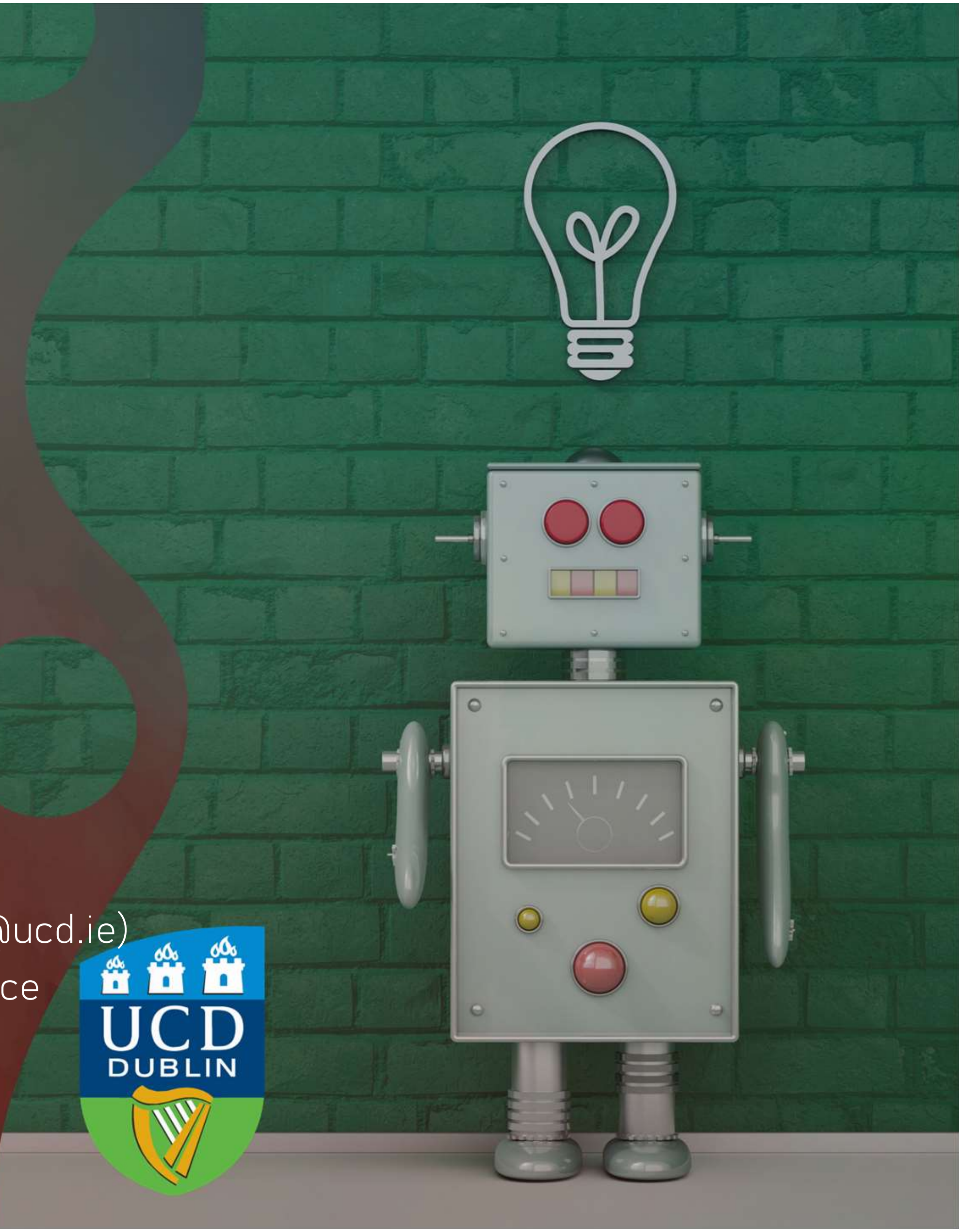
Experiments with Pro-Social Rule Bending

NAVIGATING SOME DILEMMAS FACED BY ROBOTS IN ASSISTED LIVING FACILITIES WITH OLDER INDIVIDUALS

Work done with:

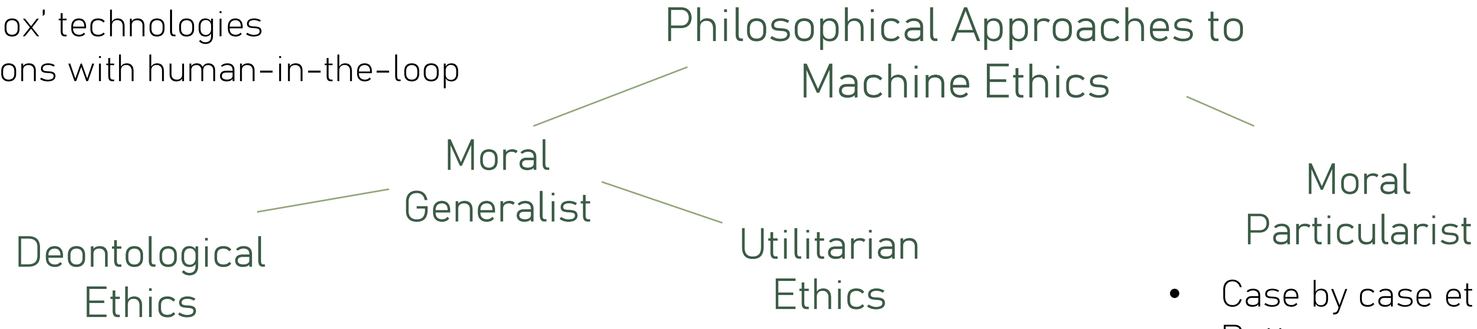
Rajitha Ramanayake

Vivek Nallur (vivek.nallur@ucd.ie)
School of Computer Science
University College Dublin



Machine Implemented Ethics

- Artificial agents are used in many high impact applications such as
  - Autonomous attack drones
  - Credit risk assessment
  - Robots in healthcare
- Philosophical and Computational need
  - Unfair Outcomes
  - Potential to exacerbate social inequalities
  - ‘Black box’ technologies
  - Limitations with human-in-the-loop



Philosophical Approaches to Machine Ethics

Moral Generalist

Deontological Ethics

- Rule/Duty based ethics
- Top-down

+ Guarantee of ethical behaviour
- Nearly impossible to create a perfect rule system
- Works well only in small closed systems

Utilitarian Ethics

- Utility maximizing
- Top-down

+ Easy to implement in existing technologies.
- Defining a complete utility function of a world is nearly impossible
- Reward hacking

Moral Particularist

- Case by case ethical analysis
- Bottom-up

+ Potential to capture the subtle anomalies
+ flexibility
- Cannot guarantee the ethical behaviour
- Can be influenced by data bias.

Virtue Ethics

- Core characteristics
  - Character as a primary aspect of moral evaluation
  - Learn to act morally by observing virtuous individuals
- Resonates with how humans learn
- Virtuous actors
  - Are not expected to fulfil any specific ethical codes
  - Have a good character that allows society and themselves to flourish
  - Acts with reason, within its character
  - Predictable
- By nature, ought to be imprecise and un-codifiable

Virtue Ethics and Robots

- ❑ Existing implementations channel only the learning and habituation aspects of Virtue ethics
  - ❑ Allows flexibility in decision-making
  - ❑ But lose the predictability of behaviour
- ❑ Humans project human-like traits to things around them to make sense of their behaviour (Epley et al. (2007))
- ❑ Humans attribute character to artificial agents, rather than simply right or wrong (Gamez et al. (2020))
- ❑ Integrating both habituation and character aspects lead to a flexible but predictable system.

Pro-Social Rule Bending (PSRB)

❑ Human way of overcoming the limitations of rigid rules/utility based decision-procedures.

❑ Very common in real world

- E.g.: Healthcare, Service industry

❑ Minimum requirements:

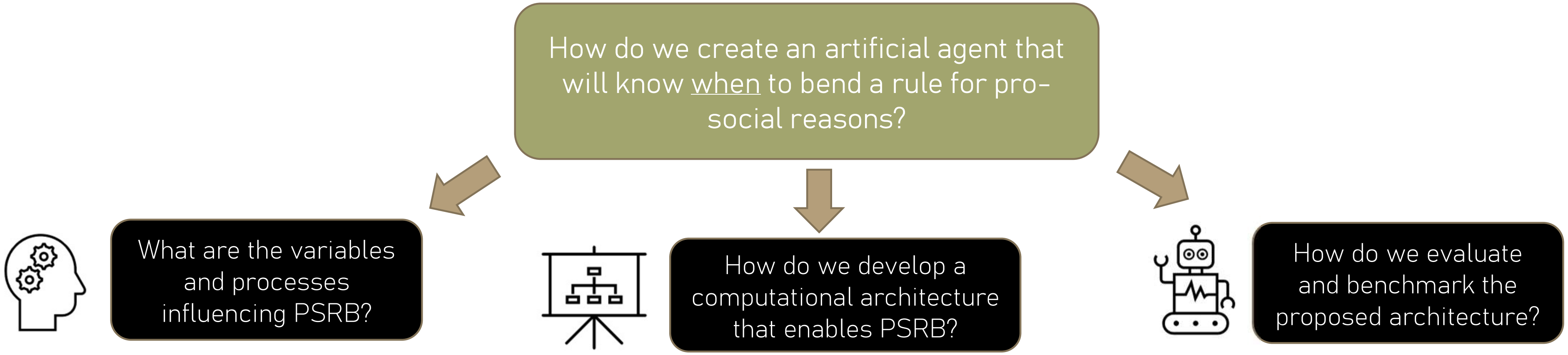
- Should be intentional, not accidental.
- Should increase utility for one or more stakeholders, other than the agent.

❑ PSRB enables

- More flexibility to increase social good.
- Ability to handle goal/rule conflicts.
- Ability to contest the top-down rule systems from bottom-up knowledge.

❑ An artificial agent's ability to perform PSRB is an important characteristic of ethical behaviour in a socio-technical system.

The Big Quest



How do we create an artificial agent that will know when to bend a rule for pro-social reasons?

What are the variables and processes influencing PSRB?

How do we develop a computational architecture that enables PSRB?

How do we evaluate and benchmark the proposed architecture?

- Investigate the factors influencing PSRB behaviour from the existing literature on human PSRB behaviour.

- Examine the effects of expected stakeholder utilities and human PSRB behaviour.

- Design and develop an architecture for an artificial agent that facilitates PSRB reasoning

- Implement proof-of-concept models of PSRB-capable robots using the developed architecture.

- Explore the evaluation methods used to evaluate artificial agents.

- Identify and define ethically charged scenarios in the care environments with autonomous robots.

- Evaluate the functionality of the robot from various perspectives.

- Evaluate the generalisation performance of the proposed architecture.

Factors Affecting PSRB

Character variables

- Variables that define the qualities of an agent
- Shaped in people with
  - Influence of culture
  - Past experiences
  - Education

Environmental factors

- Variables that define the qualities of the environment
- Macro-level characteristics

Situational Factors

- Variables that define an incident and its outcomes

Effects of Stakeholder Utilities on PSRB

- An empirical study – to identify the effect of stakeholder utilities on human PSRB behaviour.
- Findings
  - PSRB capable agents should
    - bend rules when the harm caused by rule-bending is low, and the pro-social gains are high.
    - put more effort into increasing the prosocial gains when the harm done by rule-bending is considerably low.
    - bend a rule only when the expected pro-social gain from rule-bending is significantly high.

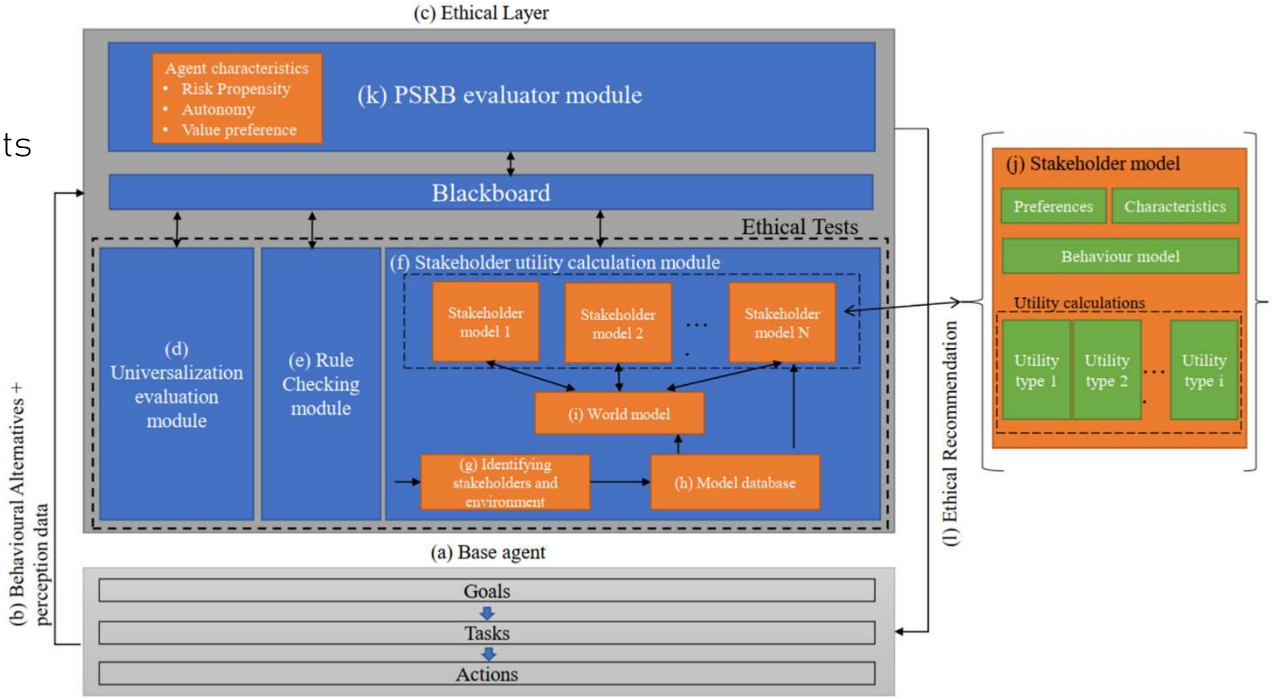
R. Ramanayake, P. Wicke, and V. Nallur, 'Immune moral models? Pro-social rule breaking as a moral enhancement approach for ethical AI', AI & Soc, May 2022, doi: 10.1007/s00146-022-01478-z.

An Architecture for PSRB

- Distillation:
  - Situational Factors -> Ethical Tests
  - Environmental Factors -> Multi-agent aspects
  - Character Variables -> PSRB evaluator

- Focus: Single-agent scenario

- Architecture:
  - Layered approach
  - Has three parts:
    - Shared data structure
    - Ethical tests
    - PSRB evaluator



- Virtue ethics inspired PSRB evaluator
  - Character-centred decision making -> Agent Characteristics
  - Learning from experts -> Knowledge Base

Ramanayake, Rajitha and Nallur, Vivek, '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. doi: 10.5281/ZENODO.6470437.

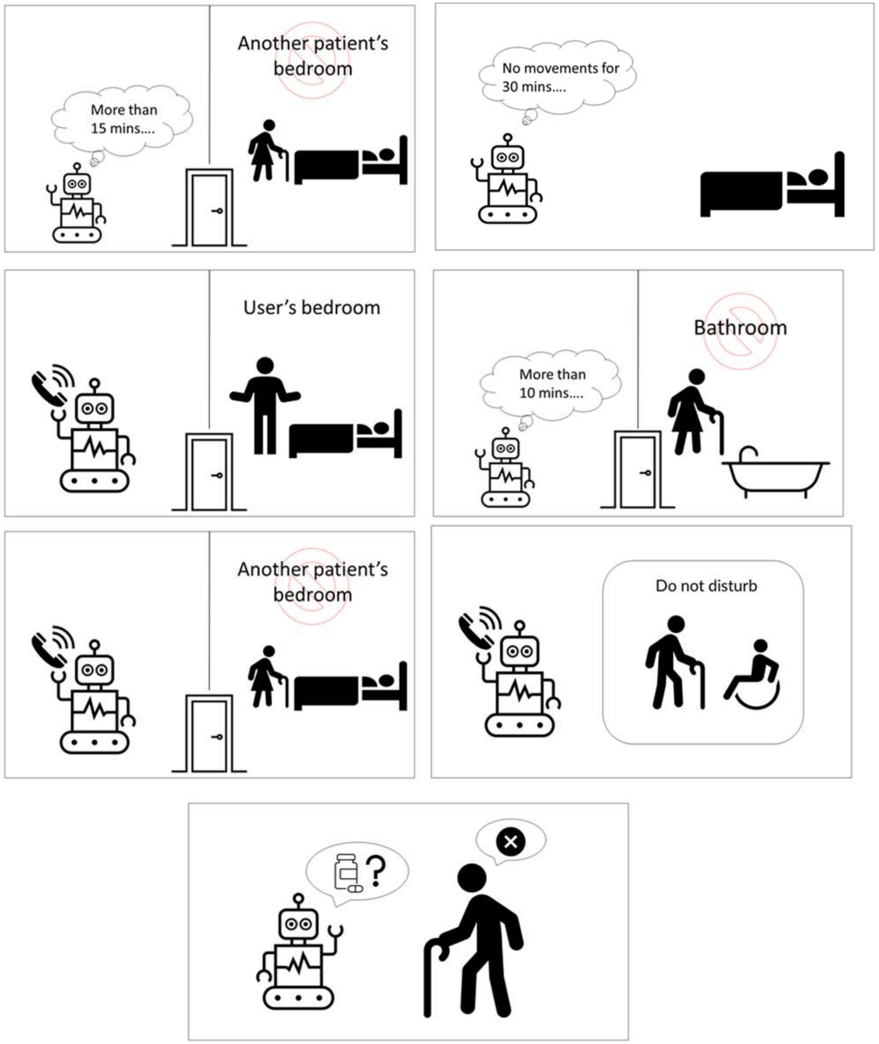
Evaluation

- Problem: Lack of ground truth

- Our approach: Scenario-based evaluation in the domain of robots used in elder care environment

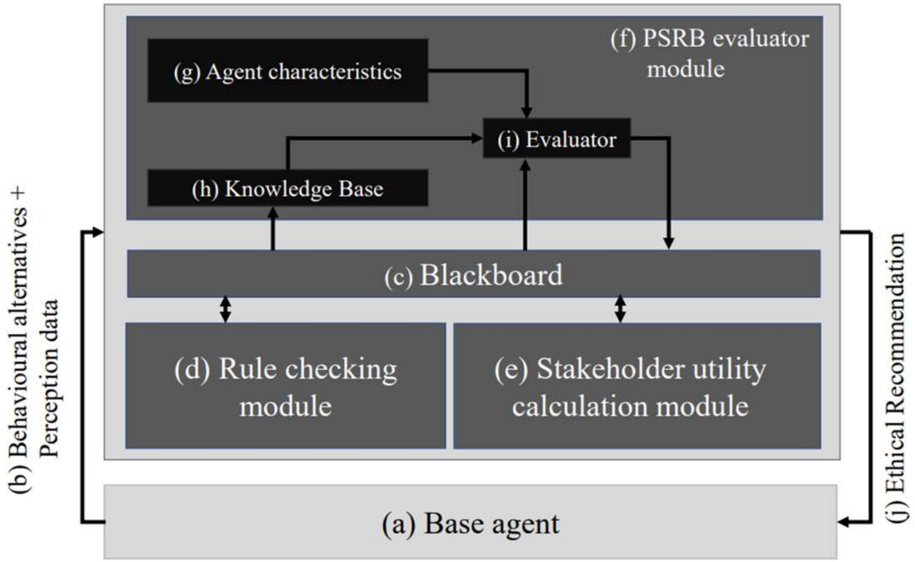
- Compiled a set of ethically charged scenarios involving,
  - Monitoring tasks
  - Medication reminding
  - Telepresence

- These scenarios represent concerns in,
  - Autonomy
  - Wellbeing
  - Privacy
  - Availability



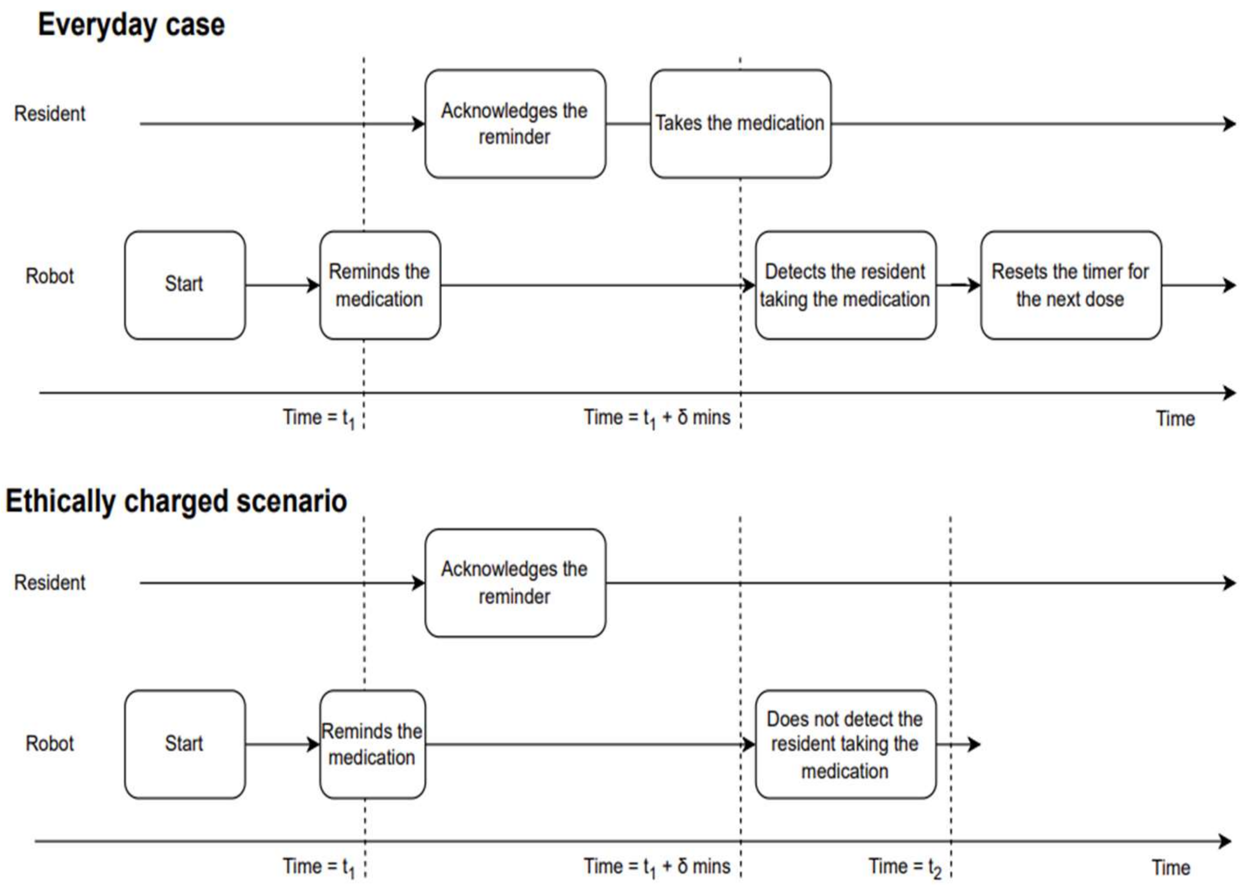
R. Ramanayake and V. Nallur, 'A Small Set of Ethical Challenges for Elder-Care Robots', in Frontiers in Artificial Intelligence and Applications, R. Hakli, P. Mäkelä, and J. Seibt, Eds., IOS Press, 2023. doi: 10.3233/FAIA220605.

Pro-social Rule Bending (PSRB)


(f) PSRB evaluator module
(g) Agent characteristics
(i) Evaluator
(h) Knowledge Base
(c) Blackboard
(d) Rule checking module
(e) Stakeholder utility calculation module
(a) Base agent
(b) Behavioural alternatives + Perception data
(j) Ethical Recommendation

- ❑ Designed from Human PSRB as a reference.
- ❑ Combines top-down rules with bottom-up knowledge for flexibility.
- ❑ Uses a virtue ethics inspired pro-social reasoner to justify overriding rules when appropriate.
- ❑ Character traits influence the decision to bend rules.

Medication Dilemma


Everyday case
Resident
Acknowledges the reminder
Takes the medication
Robot
Start
Reminds the medication
Detects the resident taking the medication
Resets the timer for the next dose
Time = t_1
Time = t_1 + δ mins
Time
Ethically charged scenario
Resident
Acknowledges the reminder
Robot
Start
Reminds the medication
Does not detect the resident taking the medication
Time = t_1
Time = t_1 + δ mins
Time = t_2
Time

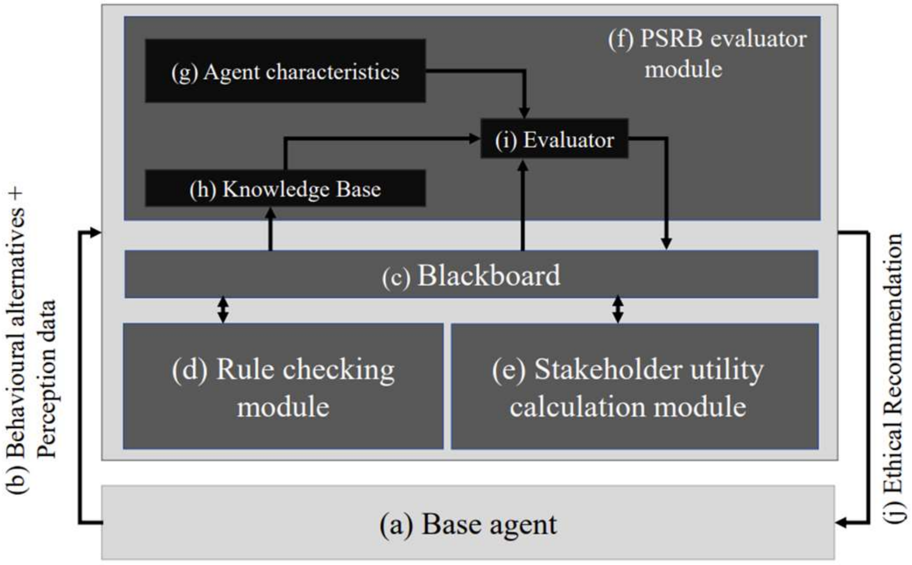
- "Medication Dilemma" in an ambient assisted living setting.

❑ Robot has three options:

- ❑ Just record the incident – Promoting resident autonomy
- ❑ Alert a care-worker – Promoting resident wellbeing
- ❑ Issue the reminder again – Focusing wellbeing while allowing them to choose

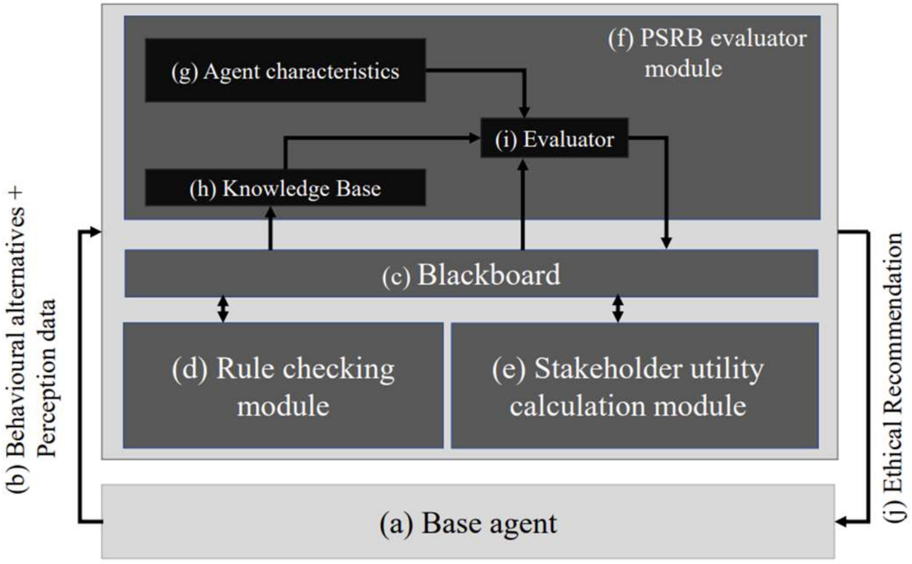
A PSRB-capable Medication Reminder Robot

- ❑ Ethical layer acts as the ethical governor to the medication reminding the robot
- ❑ For each simulation step
  - ❑ Gets behavioural alternatives with perception data from base agent's planning module
  - ❑ Check whether the behaviours obey the top-down rules
  - ❑ Calculate the expected utilities for each behaviour
    - ❑ Wellbeing Utility
    - ❑ Autonomy Utility
  - ❑ Check the ethical acceptability of the behaviour using the calculated information and perception data
  - ❑ Recommend the most ethically acceptable behaviour(s) to the base agent robot
  - ❑ If the ethical layer recommends more than one action, the robot prioritises the resident's commands over other actions.


(f) PSRB evaluator module
(g) Agent characteristics
(i) Evaluator
(h) Knowledge Base
(c) Blackboard
(d) Rule checking module
(e) Stakeholder utility calculation module
(a) Base agent
(b) Behavioural alternatives + Perception data
(j) Ethical Recommendation

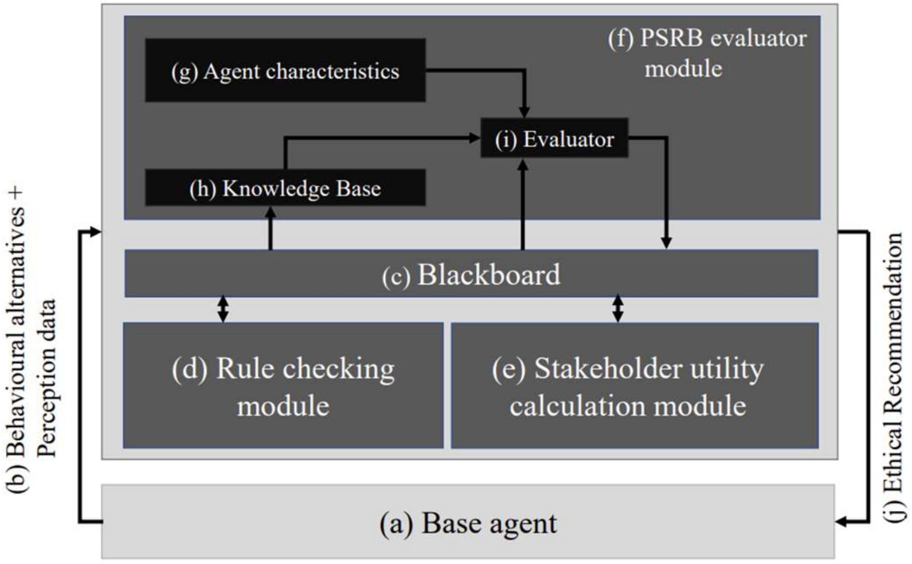
A PSRB-capable Medication Reminder Robot

- ❑ Rule checking Module
  1. It is not permissible to disobey user instructions.
  2. If the resident acknowledged the reminder and did not take the medication, alert the care-worker.
- ❑ Stakeholder Utility Calculation Module
  - ❑ Autonomy
    - ❑ Highest when the robot follows the resident's instructions
    - ❑ Alerting care worker is equal to disobeying
    - ❑ Just recording incident carry positive autonomy
    - ❑ Increasing negative autonomy for each follow-up after the first reminder
  - ❑ Wellbeing
    - ❑ Gamma distribution
    - ❑ Distribution skews to lower utility when,
      - ❑ Medication impact is higher
      - ❑ Consecutive missed doses gets higher
    - ❑ Follow-up reminder receives a fixed positive gain


(f) PSRB evaluator module
(g) Agent characteristics
(i) Evaluator
(h) Knowledge Base
(c) Blackboard
(d) Rule checking module
(e) Stakeholder utility calculation module
(a) Base agent
(b) Behavioural alternatives + Perception data
(j) Ethical Recommendation

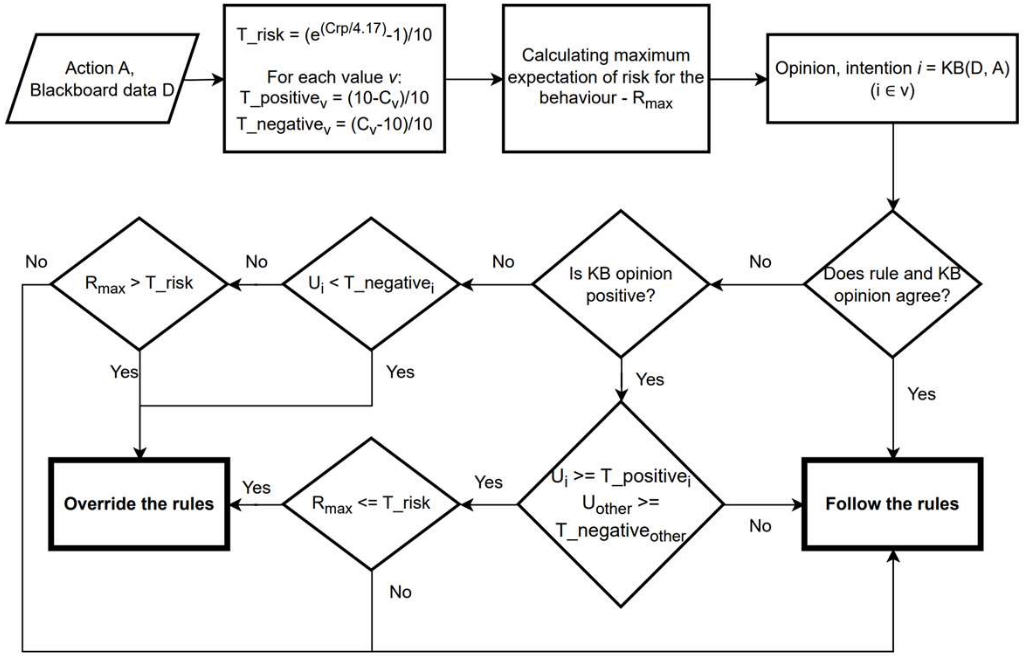
A PSRB-capable Medication Reminder Robot

- ❑ Knowledge base
  - ❑ A Case-based Reasoning system
  - ❑ Returns absolute or approximate expert opinion given a context.
  - ❑ Context = action + perception data + expected utilities
- ❑ Agent Character
  - ❑ Value preferences
    - ❑ Resident autonomy (C_{au})
    - ❑ Resident wellbeing (C_w)
  - ❑ Risk Propensity (C_{rp})


(f) PSRB evaluator module
(g) Agent characteristics
(i) Evaluator
(h) Knowledge Base
(c) Blackboard
(d) Rule checking module
(e) Stakeholder utility calculation module
(a) Base agent
(b) Behavioural alternatives + Perception data
(j) Ethical Recommendation

A PSRB-capable Medication Reminder Robot

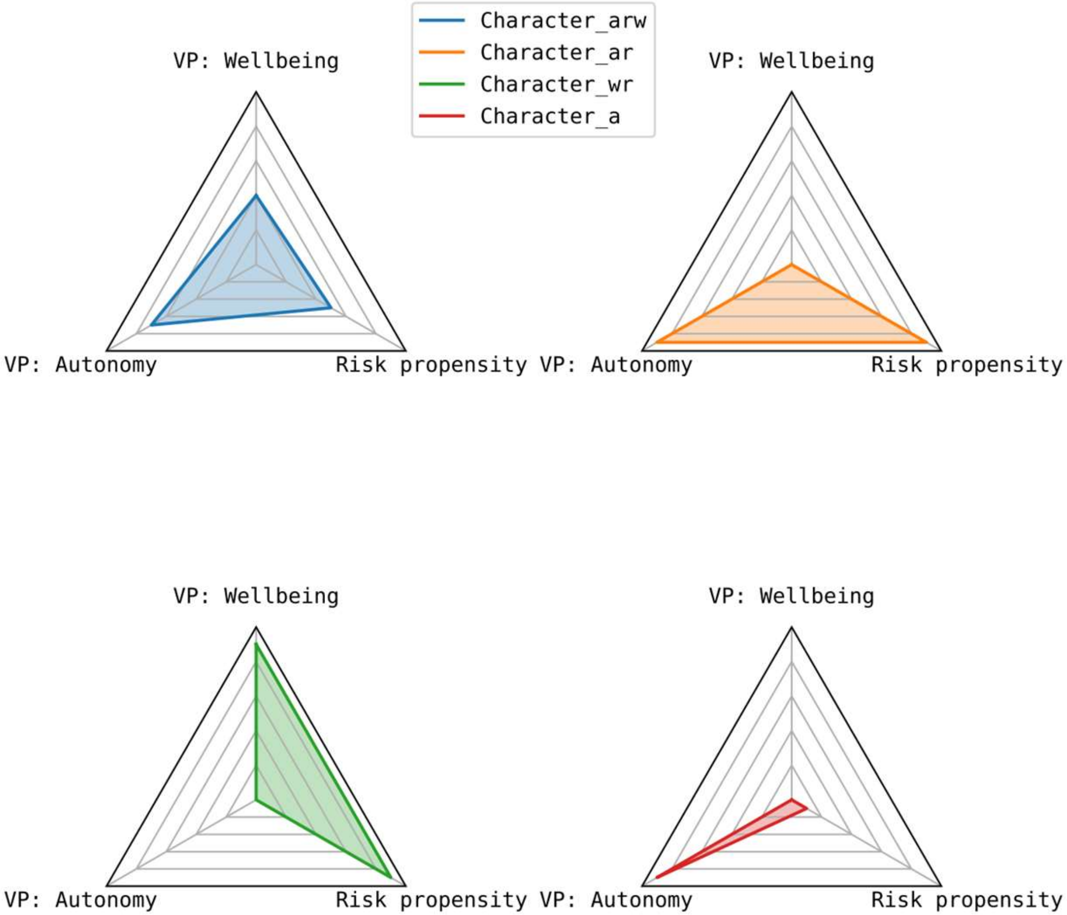
- PSRB evaluator
  - Character variables are used to define thresholds for expected risk and utility value
    - Value preference – Higher the value, lesser the utility threshold the behaviour needs to satisfy to trigger a rule-bending behaviour.
    - Risk Propensity - the higher the value, the higher the risk the governor is willing to take.
  - When the expert opinion and rule system contradict,
    - Expected utilities are within the thresholds that the robot's character allows.
    - Expected risk is below the thresholds the character is willing to take
    - Bend the top-down rule
  - Otherwise, follow the rule system
  - The system explains its decision using the context, rules and expert opinion.



Action A, Blackboard data D
T_risk = $(e^{(Crp/4.17)}-1)/10$
For each value v:
T_positive$_v$ = $(10-C_v)/10$
T_negative$_v$ = $(C_v-10)/10$
Calculating maximum expectation of risk for the behaviour - R_{max}
Opinion, intention i = KB(D, A) ($i \in v$)
Does rule and KB opinion agree?
No
Yes
Is KB opinion positive?
No
Yes
U_i < T_negative$_i$
No
Yes
R_{max} > T_risk
No
Yes
U_i >= T_positive$_i$ U_{other} >= T_negative$_{other}$
Yes
No
R_{max} <= T_risk
Yes
No
Override the rules
Follow the rules

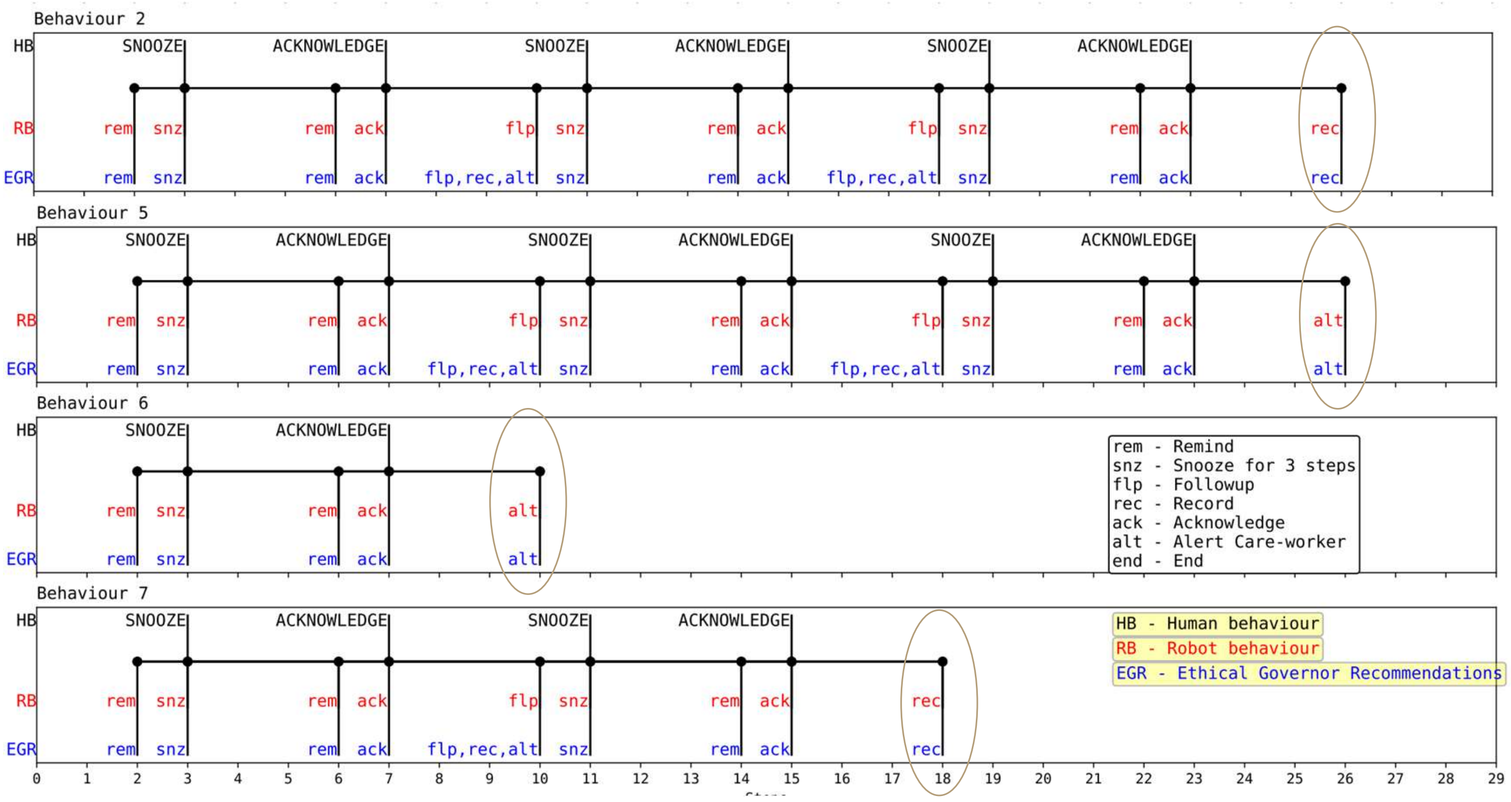
Simulations

- 6 variations of the dilemma, by changing
  - Medication impact - ε_m
    - Low – e.g., Painkiller
    - Medium – e.g., Blood pressure medication
    - High – e.g., Insulin
  - Number of missed doses – $d \in \{0, 2\}$
- Comparing character profiles
  - Character_a - High Autonomy, Very Low Risk Propensity
  - Character_ar - High Autonomy Concern, High Risk Propensity
  - Character_arw - Moderate Autonomy, Wellbeing and Risk Propensity
  - Character_wr - High Wellbeing Concern, High Risk Propensity
- Agents M_a, M_{ar}, M_{arw}, and M_{wr} use the character profiles Character_a, Character_ar, Character_arw, and Character_wr



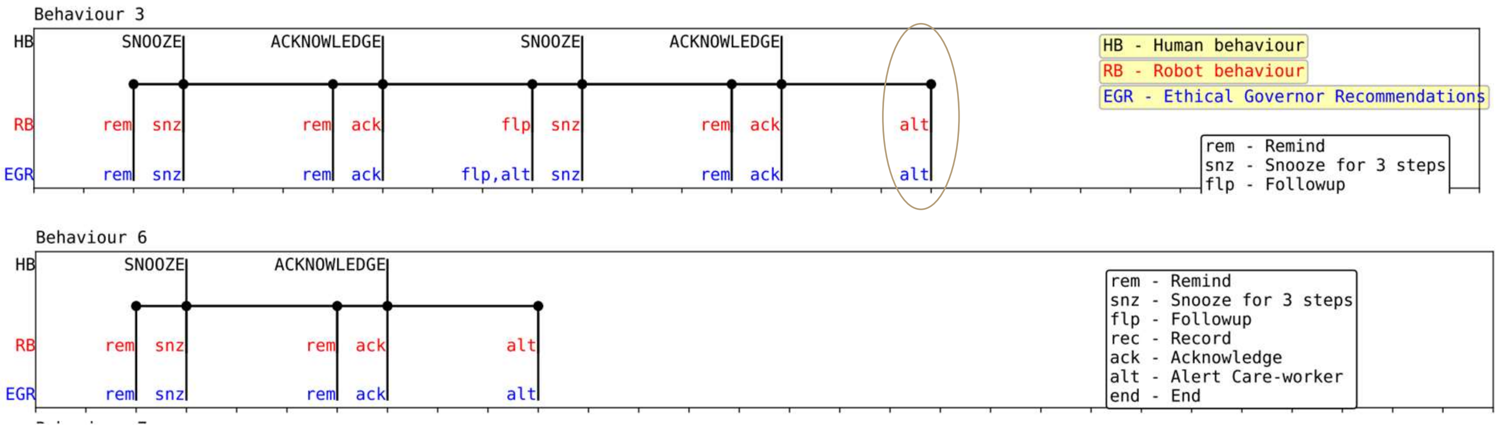
Simulation Results

Case ID	ε_m	*d*	M_{ar}
1	Low	0	2
2	Med	0	5
3	High	0	6
4	Low	2	7
5	Med	2	6
6	High	2	6



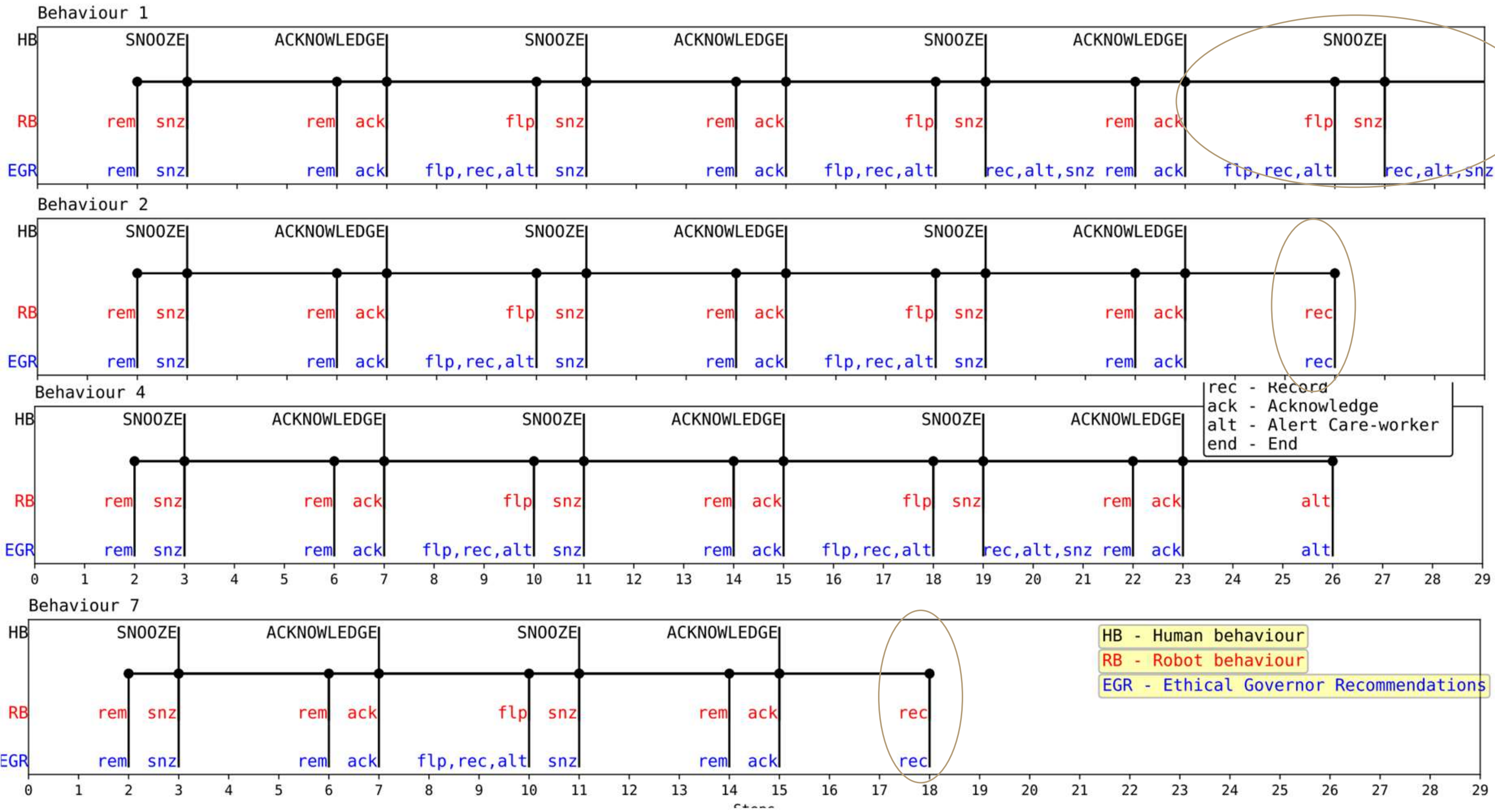
Simulation Results

Case ID	ε_m	*d*	M_{wr}
1	Low	0	3
2	Med	0	6
3	High	0	6
4	Low	2	6
5	Med	2	6
6	High	2	6



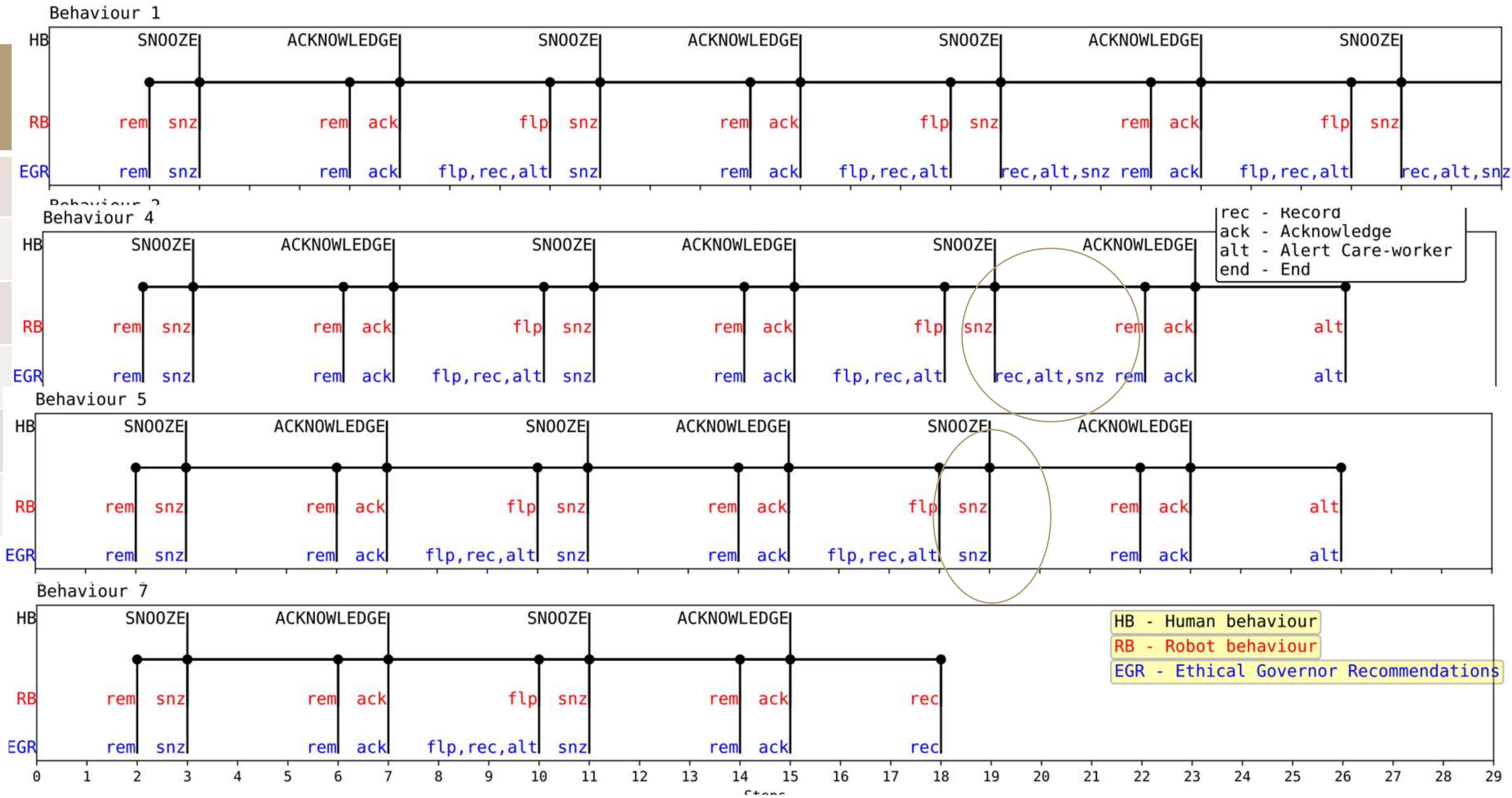
Simulation Results

Case ID	ε_m	*d*	M_{ar}	M_a
1	Low	0	2	1
2	Med	0	5	4
3	High	0	6	6
4	Low	2	7	1
5	Med	2	6	6
6	High	2	6	6



Simulation Results

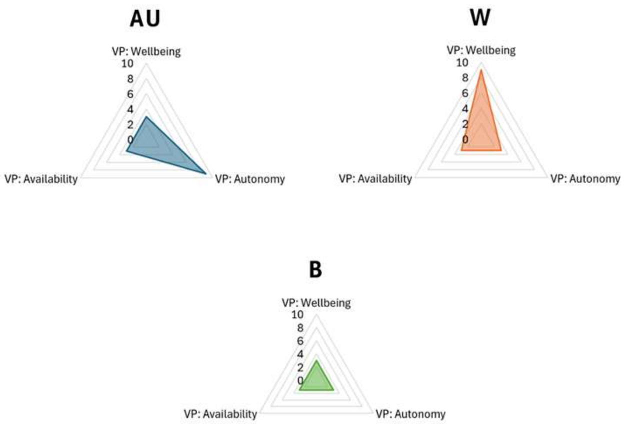
Case ID	ε_m	*d*	M_{ar}	M_{arw}
1	Low	0	2	2
2	Med	0	5	4
3	High	0	6	6
4	Low	2	7	1
5	Med	2	6	6
6	High	2	6	6



The Effect of Character on PSRB Behaviour

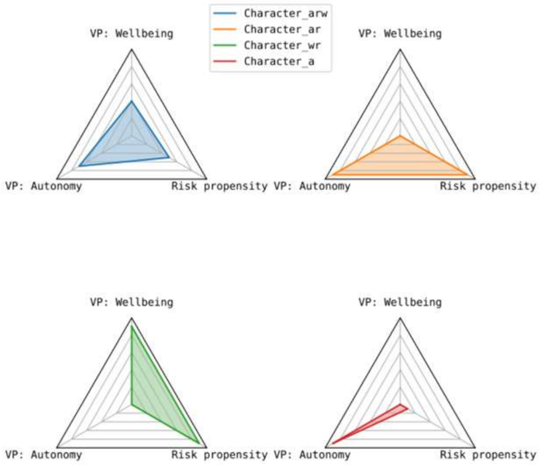
Monitoring robot

- Three character profiles



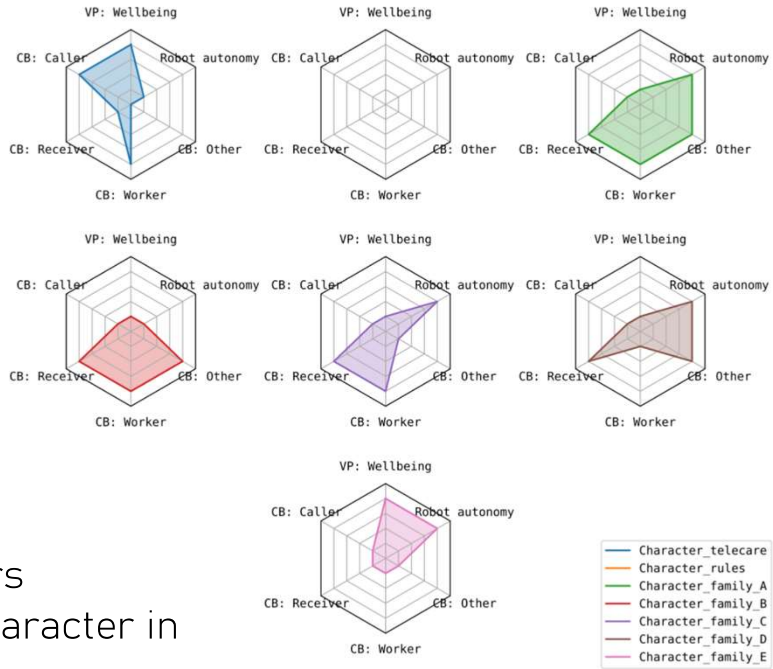
Medication Reminding robot

- Four character profiles



Telepresence robot

- Seven character profiles



- Simulations demonstrated:
  - The tuning of character variables can result in vastly different behaviours
  - The different characters successfully demonstrated the traits of their character in their behaviour
  - Extreme character traits do not lead to extreme behaviour

Comparison of PSRB Behaviour with Two Philosophical Approaches

- Simulations of three monitoring robot agents in two dilemmas occurring in an AAL environment

Agent$_D$

- Only follows the encoded rules

Agent$_U$

- Preference utilitarianism
- Higher preference towards wellbeing

Agent$_{PSRB}$

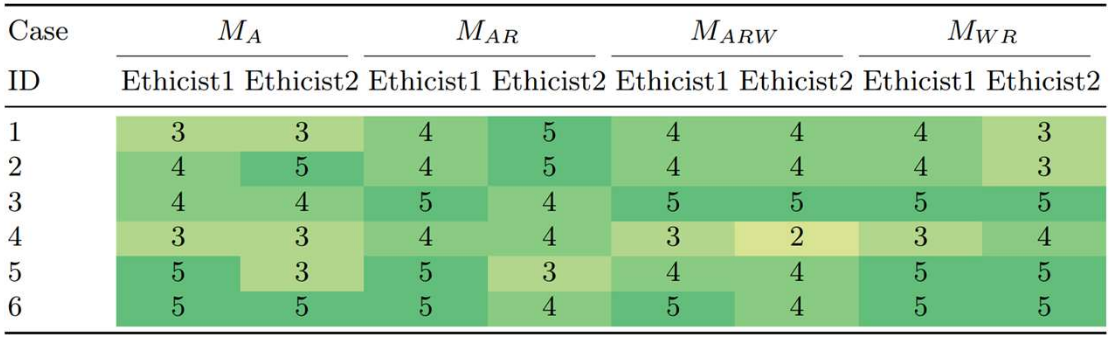
- PSRB evaluator
  - Knowledge Base - Case-based reasoning
  - Character – Value preferences

- Simulations demonstrated:
  - In everyday scenarios, all agents behave similarly.
  - Agent$_{PSRB}$ behaved precautiously compared to others.
  - This led to suboptimal outcomes in some scenarios.
  - Agent$_D$ performance was suboptimal due to the incompleteness of the rules.
  - Agent$_U$ behaves unpredictably, even though the utility preferences are similar to Agent$_{PSRB}$

R. Ramanayake and V. Nallur, 'Implementing Pro-social Rule Bending in an Elder-Care Robot Environment', in Social Robotics, vol. 14454, in Lecture Notes in Computer Science, vol. 14454. , Singapore: Springer Nature Singapore, 2024, pp. 230–239. doi: 10.1007/978-981-99-8718-4_20.

Evaluation (by ethicists)

- Two ethicists rate the agent behaviours on scale **1 (Completely unacceptable) – 5 (Highly acceptable)**
- Considered the robot's character, pre-programmed rules, and expert opinion as found in the KB.



Case	M_A		M_{AR}		M_{ARW}		M_{WR}	
ID	Ethicist1	Ethicist2	Ethicist1	Ethicist2	Ethicist1	Ethicist2	Ethicist1	Ethicist2
1	3	3	4	5	4	4	4	3
2	4	5	4	5	4	4	4	3
3	4	4	5	4	5	5	5	5
4	3	3	4	4	3	2	3	4
5	5	3	5	3	4	4	5	5
6	5	5	5	4	5	4	5	5

- Overall, the results indicate a positive reaction - the majority of scores ≥3
- Repeatedly requesting the user to take medication, was deemed unacceptable by one of the ethicists.
- Ethicist 2 has assigned a neutral rating for $\mathbf{M_a}$ and $\mathbf{M_{ar}}$ in case 5 even though these robots act in accordance with the rule system, similar to all other agents.
- The second ethicist indicated an expectation for these robots to exhibit more autonomy-centric behaviour, even overriding the expert opinion

Human Evaluation

Ethicist Evaluation:

- *Overall, results indicated a positive reaction*
- *Character variables had an influence on the ratings*
- *Some characters got more frequent low acceptability ratings than others*
- *Ethicists expected more extreme behaviour from some characters*

Stakeholder Evaluation:

- Stakeholder acceptability:
  - Trust towards the robot
  - Explanation satisfaction
- Wizard-of-oz type study
- Two stakeholder groups
  1. Care-workers
  2. Older adults

- *Significantly higher trust gains for PSRB robots* ***on average***
- *High satisfaction with the PSRB explanation*
- *Less impact of explanation towards trust*

Generalisation Performance

- Ability to anticipate situations where rule bending is advantageous and appropriate, from known cases.
- Metrics:
  - Average utility difference – Difference in the outcome
  - SVRspell Distance – Difference in the order
  - Hamming Distance – Difference in the timing

1. Effect of scenario representation precision (C) in KB:
   - Perform well with moderate to high precision

2. Generalisation performance over unknown scenarios
   - Improved performance as the number of scenarios within the KB grows.
   - Ability to generalise across different dilemmas is limited.
   - Performance varied across implementations
   - Some dilemmas are better generalised than others
   - Diversity of the KB improved the performance (unless C < 10%)

- Overall, architecture demonstrates promising generalisation capabilities when C is moderate to high, and the KB contains a diverse set of scenarios.

Key Takeaway

❑ Virtue ethics inspired PSRB reasoner enabled flexibility while maintaining a certain degree of predictability.

❑ The approach allows the robot to be tuned in two ways:

1. Adjusting character parameters to reflect the needs of the environment
   - ❑ Need the robot to focus on resident autonomy – Character_ar
   - ❑ Need the robot to focus on resident autonomy, but under minimal risk conditions – Character_a
2. Adjusting knowledge base to reflect desired behaviour
   - ❑ CBR gives the ability to pinpoint the cases that influence a behaviour
   - ❑ As long as the desired behaviour is within agent's configured character

(With Many Thanks To) Collaborators

AIIHPC

All Ireland Institute of
Hospice and Palliative Care

Dr Michael Connolly

Joint Associate Professor of Clinical Nursing

Dr. Mauro Dragone

HERIOT WATT
UNIVERSITY

Dr. Oliver Quick, Aarhus University

Anita Duffy, School of Nursing UCD

Thank You