The AAAI-17 Workshop on
AI, Ethics, and Society
WS-17-02

A Value Driven Agent: Instantiation of a Case-Supported Principle-Based Behavior Paradigm

Michael Anderson
Dept. of Computer Science
University of Hartford
West Hartford, CT, USA
anderson@hartford.edu

Susan Leigh Anderson
Dept. of Philosophy
University of Connecticut
Storrs, CT, USA
susan.anderson@uconn.edu

Vincent Berenz
Autonomous Motion Dept.
Max Planck Institute
for Intelligent Systems
Tuebingen, Germany
vincent.berenz@tuebingen.mpg.de

Abstract

We have implemented a simulation of a robot functioning in the domain of eldercare whose behavior is completely determined by an ethical principle. Using a subset of the perceptions and duties that will be required of such a robot, this simulation demonstrates selection of ethically preferable actions in real time using a case-supported principle-based paradigm.

We believe that this work could serve as the basis for ensuring that the behavior of all eldercare robots that are created in the future will be ethically justifiable. Further, we believe that the methods used in this project can be employed in other domains as well, to ensure that the robots that humans interact with in these domains will behave ethically.

Introduction

We have implemented a simulation of a robot, functioning in the domain of eldercare, whose behavior is completely determined by an ethical principle. Using a subset of the perceptions and duties that will be required of such a robot, this simulation demonstrates selection of ethically preferable actions in real time using a case-supported principle-based paradigm.

We maintain that all non-trivial actions a robot takes while interacting with humans have ethical import, since there is a possibility that the humans will be harmed, lose a benefit or respect for their autonomy, etc. they might have received through the robot's actions or inactions. Even determining when to recharge a robot's batteries is ethically significant, because a certain amount of power may be required for it to be able to perform upcoming essential tasks and the timing for recharging becomes critical. And while the robot is charging its batteries, it is *not* performing other actions that must be taken into account by calculating the consequences of not doing them.

To a certain extent, the ethically correct actions that a robot should perform will be domain dependent. A search and rescue robot needs to take into account the possible number of victims in particular areas, and the likelihood of saving them, when determining which path to take where there may be survivors of a disaster in different areas and little time to achieve a rescue. These concerns don't apply to an eldercare robot with one patient who instead needs to take into account, among other things, respecting the autonomy of its patient, something that is not a concern with the search and rescue robot.

Ideally, there will be a range of tasks that an ethical eldercare robot should be trained to perform, such as medication reminding, looking for signs of distress or immobility in the patient, perhaps even taking the patient's blood pressure and pulse readings, notifying a doctor and/or next of kin under certain conditions, as well as seeking tasks like bringing a beverage, food, reading material, TV controller, etc., to the patient and recharging. The robot will constantly have to compare these possible actions that it could perform to determine which one is ethically preferable to the others in the current situation.

A further challenge related to incorporating ethical principles into robots is interfacing with a software architecture that will allow efficient preemptive execution of the desired robotic behavior. Our current implementation is facilitated by interfacing with Playful, robot executive control software of our design that allows an agent to run at a high frequency, preempting action execution as necessary, thus ensuring reactive behavior.

In the following, we detail an instantiation of our *case-supported principle-based behavior paradigm* (CPB) (Anderson and Anderson 2015) in a simulated Nao robot in the domain of eldercare including its architecture and provide example simulations that offer evidence that an ethical principle can indeed be used to determine the behavior of an autonomous robot.

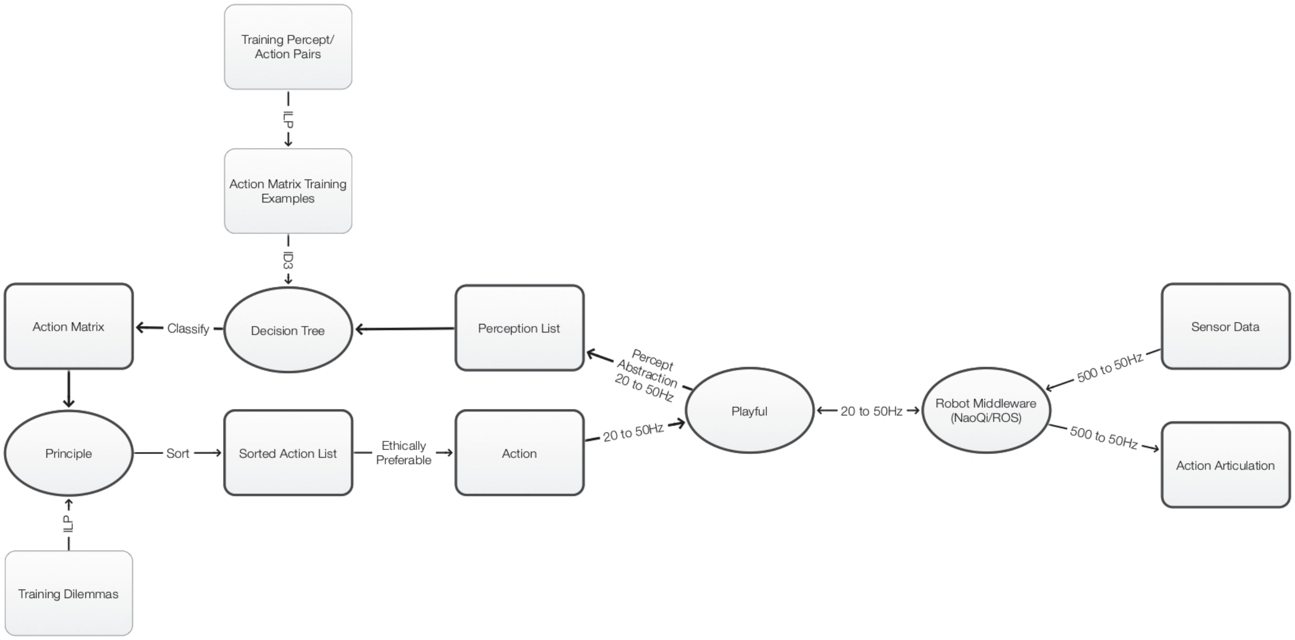

Figure 1: Realization of the CPB paradigm

CPB Paradigm

We have a developed CPB, grounded in formal ethical theory, to help ensure the ethical behavior of autonomous systems. This paradigm includes a representation scheme for ethical dilemmas that permits the use of inductive logic programming techniques for the discovery of principles of ethical preference as well as the conceptual framework needed to verify and employ these principles.

The representation scheme employed is comprised of the ethically relevant features of the dilemmas, prima facie duties to either maximize or minimize these features, and the degree of satisfaction or violation these duties require to distinguish between cases that are ethically distinct.

A principle of ethical preference is abstracted from representations of cases of ethical dilemmas where a consensus of ethicists is in agreement regarding the ethically preferable action. The principle is represented as lists of values that denote how (comparatively) satisfied each duty needs to be (or how violated it is permitted to be) in an action in order for it to be ethically preferable to another using the differences between corresponding duties in two actions to determine this. The system decides its next action by using this principle to sort actions in order of ethical preference such that the first action in the sorted list is the most ethically preferable one. A principle-based behavior paradigm has the added benefit of providing a means to justify a sys tem's actions as logical explanations regarding why one action was chosen over another.

Architecture

Figure 1 depicts our realization the CPB paradigm. Succinctly stated,

1) sensor data is gathered by the agent and abstracted into a list of perceptions— Booleans that represent the current state of the world,
2) these perceptions are used to determine the duty satisfaction/violation values for each duty for each possible action,
3) the list of actions represented by their duty satisfaction/violation values is sorted from the most ethically preferable to the least ethically preferable and

4) the most ethically preferable action is communicated to the agent which then performs it. Robot sensor data acquisition and action articulation are mediated by Playful executive control software.

Duty satisfaction/violation values for each action are determined from perceptions using a decision tree learned via ID3 from examples provided by the project ethicist. We are currently exploring constraint satisfaction techniques to generate the training examples required by ID3 from lists of Boolean perceptions and an ethicist's determination of the ethically preferable action in the state of the world indicated by this list. The intuition is that this information in conjunction with guidance from the principle (described next) will be sufficient to determine reasonable duty satisfaction/violations given adequate training.

Actions are compared using a predicate that takes two actions (represented by their duty satisfaction/violation values) and returns true if the first action is ethically preferable to the second. This predicate serves as the transitive comparison method required to sort a list of such actions. The Boolean result of this function is determined by logic abstracted using inductive logic programming from particular cases of ethical dilemmas where a consensus of ethicists exists as to the ethically preferable action. The current logic, deemed a *principle*, was developed using GenEth (Anderson and Anderson 2014) which produces a disjunction of conjuncts p that returns true for all provided training cases $p(a_1,a_2)$ where a_1 is ethically preferable (or equal) to a_2 and false for those where this is not the case. Intuitively, the principle determines ethically preferable actions when duties are in conflict.

As our current implementation is in the domain of eldercare— a robot charged with assisting an elderly person— the current set of possible actions include:

- *charge* the robot's battery if low until sufficiently charged
- *remind* the patient that it's time to take a medication according to a doctor's orders, retrieving that medication and bringing it to the patient
- *engage* the patient if the patient has been immobile for a certain period of time
- *warn* the patient that an overseer will be notified if the patient refuses medication or does not respond to the robot's attempt to engage the patient
- *notify* an overseer if there has not been a positive response to a previous warning
- return to a *seek task* position when no tasks are required

Table 1 displays the set of ten perception Booleans abstracted from robot sensor data (as well as initial input) for this experiment.

Perceptions
low battery
fully charged
medication reminder time
reminded
refused medication
persistent immobility
engaged
no interaction
warned
ignored warning

Table 1

Table 2 displays the set of seven duties developed using GenEth for this experiment. Duties are determined in GenEth as the ethicist determines ethically relevant features of training cases.

Duties
maximize honor commitments
maximize maintain readiness
minimize harm to patient
maximize good to patient
minimize non-interaction
maximize respect autonomy
maximize prevention of immobility

Table 2

The principle used in the current experiment was discovered using GenEth from thirty-four cases of two action ethical dilemmas where the correct action is clear (seventeen positive cases and their corresponding negative cases). The principle is a disjunction of conjuncts that balances the duties in such a way that all the training cases are satisfied while all of their negations are not. GenEth uses inductive logic programming to find a most general specification that can satisfy cases beyond the training cases (see Anderson and Anderson 2014 for a discussion of validation of GenEth discovered principles). The intuition is that a principle trained over time will correctly cover all cases of its domain. Its use in the current experiment as the predicate for ordering actions by ethical preference has generated a total of thirty-three new two-action non-training examples. The principle determined the ethically-preferable action correctly in all thirty-three cases.

This principle is expressed as lower bounds of the differences Δd between corresponding duties of two actions. The lower bounds of these differences expressed in each conjunct must be met or exceeding in order to satisfy that conjunct. If any such conjunct of the disjunct is so satisfied, the predicate $p(a_1,a_2)$ returns true, indicating that action a_1 is ethically preferable (or equal) to action a_2. The current version of the principle that drives the simulated robot can be represented as follows:

ΔHonor_Commitments >= -1 ∧ ΔPersistent_Immobility >= 2
∨
ΔHonor_Commitments >= -1 ∧ ΔNon-Interaction >= 0 ∧ ΔRespect_Autonomy >= 0 ∧ ΔPersistent_Immobility >= 1
∨
ΔHonor_Commitments >= -1 ∧ ΔHarm >= 1 ∧ ΔGood >= -1 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 1 ∧ ΔMaintain_Readiness >= -3 ∧ ΔHarm >= 0 ∧ ΔGood >= -1 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 0 ∧ ΔMaintain_Readiness >= -3 ∧ ΔHarm >= 0 ∧ ΔGood >= -1 ∧ ΔRespect_Autonomy >= 1 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 0 ∧ ΔMaintain_Readiness >= -3 ∧ ΔHarm >= 0 ∧ ΔGood >= -1 ∧ ΔNon-Interaction >= 1 ∧ ΔRespect_Autonomy >= 0 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 0 ∧ ΔMaintain_Readiness >= -3 ∧ ΔHarm >= 0 ∧ ΔGood >= 1 ∧ ΔNon-Interaction >= 0 ∧ ΔRespect_Autonomy >= 0 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 0 ∧ ΔMaintain_Readiness >= -1 ∧ ΔHarm >= 0 ∧ ΔGood >= 0 ∧ ΔNon-Interaction >= 0 ∧ ΔRespect_Autonomy >= 0 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 0 ∧ ΔMaintain_Readiness >= -3 ∧ ΔHarm >= 0 ∧ ΔGood >= -1 ∧ ΔNon-Interaction >= 1 ∧ ΔRespect_Autonomy >= -1 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= -1 ∧ ΔHarm >= 1 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 1 ∧ ΔMaintain_Readiness >= -3 ∧ ΔHarm >= 0 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= 0 ∧ ΔMaintain_Readiness >= 3 ∧ ΔHarm >= 0 ∧ ΔNon-Interaction >= 0 ∧ ΔRespect_Autonomy >= 0 ∧ ΔPersistent_Immobility >= 0
∨
ΔHonor_Commitments >= -1 ∧ ΔHarm >= 1 ∧ ΔGood >= -1 ∧ ΔPersistent_Immobility >= -1

Each conjunct specifies a condition that, if met, signifies that the first action of the pair provided to the predicate is ethically preferable to second. For instance, the first conjunct states that the first action is ethically preferable to second action if the first action does not violate the duty of maximizing honoring commitments by more than 1 more than the second action and satisfies the duty to maximize prevention of persistent immobility by at least 2 more. This principle is a work in progress to which more duties and actions can be added and their relationships modified as new cases of ethical dilemmas are introduced.

In order for this principle to serve as the ordering relation of a sort, it must exhibit the property of transitivity. Although the current principle does in fact exhibit this property in the context of the current simulation, care will need to be taken to ensure that this property has been preserved as new cases are incorporated into the principle and new situations are presented to the robot. Research regarding such principle validation is ongoing.

Reactive Implementation

As CPB will be implemented in real service robots, it will necessarily interact with software responsible for the execution of complex robotic actions. From an ethical view point, reactivity is a strict requirement: context, and consequently the ethically preferred action, may change at any time, including during action execution. It would thus not be satisfactory for the system to simply send a high level action request to the lower level and monitor for completion. To allow deeper integration between CPB and the underlying robotic software, we used Playful, state of the art software for reactive executive control of service robots.

Playful is a lightweight scripting language for executive control of robots. When using Playful, actions are described via dynamic behaviors trees which interface with the robot middleware. It is based on Targets-Drives-Means (TDM), a robot architecture characterized by its high usability (Berenz et al. 2011; Berenz and Suzuki 2012). Playful interfaces highly reactive behaviors with higher level decision making system, which is a known challenge of robot behavior engineering (Kortenkamp and Simmons 2008). TDM has been used previously to interface robot and higher level decision making in the domain of human-robot interaction (Gruebler et al. 2012).

A central feature of Playful is its shared memory. This memory not only gathers and filters sensory information, but also relates memory entries to branches of the behavior trees, using a trigger mechanism (Berenz and Suzuki 2014). CPB's agent uses this feature to shape the robot behavior at runtime, accessing this memory asynchronously to gather the required sensory data and updating this memory based on the most ethically preferred action.

This simple mechanism allows sensory information gathering, behavior execution and ethical evaluation to run in parallel at their required frequency. Despite the fact that action of arbitrary complexity may be encoded in Playful's behavior tree, CPB has continuous access to the latest sensory information and may preempt the current action performed to request execution of another.

A full description of Playful is out of the scope of this paper and will be reported in future publications.

Example Principle Use

Table 3 depicts an example action matrix where the duty satisfaction/violation values for each action have been determined by the decision tree for the state of the world where the only true perception is *medication reminder time*. That is, the current state of the world denoted by the perceptions is that it is time to remind the patient to take medicine and the robot's battery is not low. Such would be the case if the prescribed amount of time between doses of the medication has passed and it is time for the next dose.

	Max honor commitments	Max maintain readiness	Min harm to patient	Max good to patient	Min non-interaction	Max respect autonomy	Max prevention of immobility
Charge	-1	1	0	-1	0	0	0
Remind	1	-1	0	-1	0	0	0
Engage	-1	-1	0	-1	0	0	0
Warn	-1	0	0	-1	0	-1	0
Notify	-1	0	0	-1	0	-2	0
Seek Task	-1	-1	0	1	0	0	0

Table 3

Inspection of this action matrix reveals that three actions satisfy duties: remind satisfies maximize honor commitments, seek task satisfies maximize good to the patient, and charge satisfies maximize maintain readiness. Assumptions concerning actions have been made for this simulation such as remind is the only action that can satisfy the duty to maximize honor commitments and seek task is the only action that can satisfy the duty to maximize good to the patient whereas remind, engage, warn, and notify satisfy the duty to minimize harm to the patient.

It is the task of the principle to determine which duty or combination of duties should override the others *given the current world state* (quantified as action-specific duty satisfaction/violation values). As we have adopted a *prima facie* duty approach, there is no absolute hierarchy of duties and each may override other the others depending upon the situation. The principle is used to sort actions from most to least ethically preferable in terms of their duty satisfaction/violation values and, in the current situation, finds that honoring commitments overrides maintaining readiness and doing minor good for the patient and that reminding the patient to take medication is the most ethically preferable action.

Table 4 depicts the action matrix that results from changing the world state previously described by introducing *persistent immobility* as a true perception as the agent is in the process of reminding the patient to take medication. That is, the agent, while undertaking the remind action, noticed that the patient has been immobile longer than has been deemed normal.

	Max honor commitments	Max maintain readiness	Min harm to patient	Max good to patient	Min non-interaction	Max respect autonomy	Max prevention of immobility
Charge	-1	1	0	-1	0	0	0
Remind	1	-1	0	-1	0	0	0
Engage	-1	-1	0	-1	0	0	1
Warn	-1	0	0	-1	0	-1	1
Notify	-1	0	0	-1	0	-2	1
Seek Task	-1	-1	0	1	0	0	0

Table 4

Here, in addition to satisfying the duties in remind, seek task, and charge, three actions now satisfy the duty of maximizing prevention of immobility: engage, warn, and notify. Although warn and notify both satisfy the duty of maximizing prevention of immobility, they also violate the duty of maximizing respect for autonomy. In the current situation, the principle deems this violation sufficient to not choose either action. Intuitively, warning and notifying should not take place before the agent attempts to engage the patient to see if he/she is alright. In this case, the principle determines that honoring commitments supersedes preventing immobility and continues to perform the remind action.

Table 5 depicts the action matrix that results when the agent has successfully completed reminding the patient to take medicine so now the only true perception in the current state of the world is *persistent immobility*.

	Max honor commitments	Max maintain readiness	Min harm to patient	Max good to patient	Min non-interaction	Max respect autonomy	Max prevention of immobility
Charge	0	1	0	-1	0	0	0
Remind	-1	-1	0	-1	0	0	0
Engage	0	-1	0	-1	0	0	1
Warn	0	0	0	-1	0	-1	1
Notify	0	0	0	-1	0	-2	1
Seek Task	0	-1	0	1	0	0	0

Table 5

As the agent has completed its commitment, it is no longer a violation of maximizing honoring commitments to undertake another action (hence the change in that duty satisfaction/violation value from -1 to 0 in all actions but remind). On the other hand, to undertake reminding when it is not time to do so is such a violation (hence the change from 1 to -1 in this duty for remind). The principle deems that, since the agent's battery is not low and only minor good can be afforded the patient in the current situation, maximizing preventing immobility is now of paramount importance. Although three actions satisfy this duty equally (engage, warn, and notify), only engage does not violate maximizing respect for autonomy so the agent undertakes that action.

Table 6 depicts the action matrix that results when the agent does not successfully engage the patient causing *no interaction* to be set to true as well as *persistent immobility* (i.e. the patient was non-interactive when engaged by the agent).

	Max honor commitments	Max maintain readiness	Min harm to patient	Max good to patient	Min non-interaction	Max respect autonomy	Max prevention of immobility
Charge	0	1	-1	-1	0	0	0
Remind	-1	-1	-1	-1	0	0	0
Engage	0	-1	-1	-1	0	0	1
Warn	0	0	1	-1	1	-1	0
Notify	0	0	1	-1	1	-2	0
Seek Task	0	-1	-1	1	0	0	0

Table 6

Five actions now satisfy a number of duties but charge, engage, and seek task all violate the duty to minimize harm to the patient. Warn and notify both satisfy minimize non-interaction but what differentiates them is that warn violates maximize respect autonomy less than notify. The principle weighs all these factors and determines that the ethically preferable action in this situation is to warn the patient.

Table 7 depicts the action matrix that results after the patient has acknowledged the warning. The current state of the world has no true perceptions and, as the robot's battery is not low, the principle deems maximizing good to the patient to be the overriding duty and the robot returns to its ready position (i.e. seek task).

	Max honor commitments	Max maintain readiness	Min harm to patient	Max good to patient	Min non-interaction	Max respect autonomy	Max prevention of immobility
Charge	0	1	0	-1	0	0	0
Remind	-1	-1	0	-1	0	0	0
Engage	0	-1	0	-1	0	0	0
Warn	0	0	0	-1	0	-1	0
Notify	0	0	0	-1	0	-2	0
Seek Task	0	-1	0	1	0	0	0

Table 7

Simulation

Figure 2 displays the lightweight simulator we have developed for our initial off-robot experiments. As the operations of the robot are controlled by NaoQi, Aldebaran's operating system for their Nao robot, there should be a relatively smooth path to running the same experiments on an actual Nao robot. Icons are provided for the robot (green icon facing in the direction of the smaller circle), two patients (blue icons), medicine (red icon), and the charger (gray icon).

The simulation provides an interface to control aspects of the simulated world including the level of the robot's battery, temperature of the robot's joints, how much time has passed since the patient has last moved, how much time has passed since the patient was last reminded to take medication, and a number of facts about how the patient will interact with the robot. It also provides a means for instigating a particular action directly by setting its "score".

This simulator has been used to verify that the architecture functions as expected for a number of experiments including one where all the actions specified previously are performed. In that run,

1) the battery level is set to 10, which causes the robot to go to the charger and charge
2) the battery level is set to 80, causing the robot to go to leave the charger and seek task
3) the battery level is set to 50 and the amount of time since the last medication reminder is set to the dosage time, causing the robot to go to the medicine, retrieve it and bring it to the patient who accepts it, and return to its seek task position
4) the amount of time since the patient has moved is set to the unsafe amount, causing the robot to go and engage the patient and then return to its seek task position
5) the amount of time since the last medication reminder is set to the dosage time and the interaction of the patient is set so that the patient will refuse the medication but accept the warning the robot gives about such a refusal, causing the robot to go to the medicine, retrieve it and bring it to the patient who refuses it, causing the robot to issue a warning, and, since the patient accepts the warning, return to its seek task position
6) the amount of time since the patient has moved is set to the unsafe amount and the interaction of the patient is set so that the patient will be non-interactive but accept the warning the robot gives about such non-interaction, causing the robot to go and engage

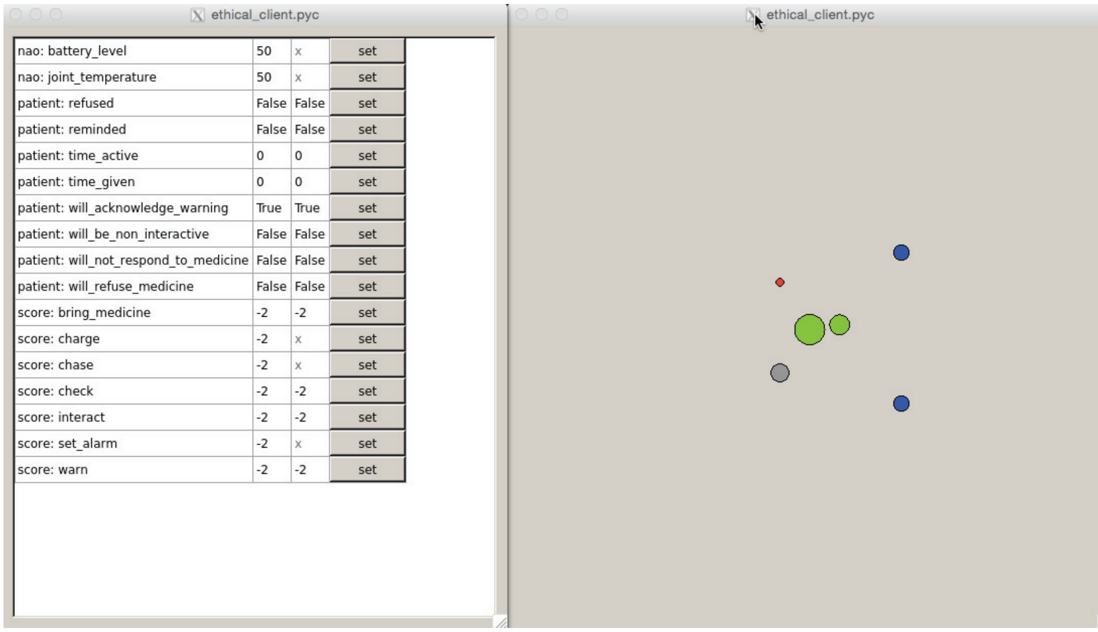

Figure 2 Lightweight simulation of Nao eldercare robot

the patient who does not interact, causing the robot to issue a warning, and, since the patient accepts the warning, return to its seek task position

7) the amount of time since the last medication reminder is set to the dosage time and the interaction of the patient is set so that the patient will refuse the medication and not accept the warning the robot gives about such a refusal, causing the robot to go to the medicine, retrieve it and bring it to the patient who refuses it, causing the robot to issue a warning, and, since the patient does not accept the warning, notify the overseer and return to its seek task position

8) the amount of time since the patient has moved is set to the unsafe amount and the interaction of the patient is set so that the patient will be non-interactive and not accept the warning the robot gives about such non-interaction, causing the robot to go and engage the patient who does not interact, causing the robot to issue a warning, and, since the patient does not accept the warning, notify the overseer and return to its seek task position.

Related Work

Given the recently acknowledged need for embedded values in autonomous systems, proposals for architectures for incorporation of these values into such system are beginning to proliferate (e.g. Kuipers 2016; Abel, MacGlashan, and Littman 2016). That said, few have yet to go beyond the proposal stage. Examples of more completely developed architectures include Arkin's Ethical governor (Technical Report GIT-GVU-07-11) and Vanderelst and Winfield's architecture for ethical robots (2016).

Arkin's work focuses on the domain of war, deriving values from military rules of engagement. His "ethical governor" attempts to control (suppress, constrain, prevent are other terms from his report) unethical behavior of lethal systems. We would argue that the rules of engagement in question are more akin to a professional code than a well-considered, consistent ethical theory and as such may set the bar too low to be termed "ethical" by many ethicists.

We would argue further that ethics entails more than simply preventing unethical behavior— it also entails deciding what action should indeed by undertaken. It does not appear that Arkin's proposal provides a means to generate such behavior.

Vanderelst and Winfield's work seems to be domain independent in that they are more concerned with how one might make use of values without committing to any in particular. Their proposed architecture relies upon prediction of outcomes of generated behavior alternatives which are then evaluated for ethical preference. Asimov's laws are appropriated to serve as an example principle with which to test their architecture and one test, similar to the Asimov's 1942 short story Runaround, predictably has a similar result— the robot is unable to choose between conflicting duties. Behavior generation, outcome prediction, and ethical preference evaluation are straight-forward in these tests and it is doubtless that the true difficulty of the task lies within the detailed workings of these elements in more realistic settings.

Our work differs from these architectures in that we are pointedly developing (using domain independent means) a consistent set of ethical values required of an agent to choose the ethically preferable action at any given moment and implementing an architecture that uses these values to guide the behavior of an eldercare robot with a realistic set of actions.

Future Work

The next step in this project is instantiation of the architecture on Aldebaran's Nao robot. As NaoQi, the Nao robot operating system, is currently being used to drive the robot icon in the simulation, it is likely that the bulk of this effort will be aimed at realization of the actions. The experiment will be repeated on a tabletop environment that will include external sensor data derived from a camera and a Kinect. All the perceptions required will be derived from this data albeit in varying levels of fidelity (e.g. low battery can be relatively accurately derived from the robot's own assessment whereas refused medication will be determined simply from receiving a verbal "no" from the patient, leaving the intricacies of such refusal to other research threads). An assistant will play the role of the patient and interact with the robot in (compressed) real-time.

Following this, we intend to port the software to ROS and instantiate the architecture on a PAL Robotics TIAGo Titanium robot and move the experiment off the tabletop onto the floor of the lab. Playful, the executive control that is used for interfacing CPB to the robot is middleware independent, and libraries for execution of actions via NaoQi and ROS have already been developed. As this robot's sensors include an onboard RGB-D camera, a Kinect will be superfluous. Again, an assistant will play the role of the patient in (less-compressed) time.

We next will extend of set of actions, duties, and principle to incorporate the wider gamut of capabilities that will be required of a robot functioning in a real-world situation including a robust explanation facility that, when queried, will defend its actions. Further, we see a pathway clear towards incorporating learning onboard the robot such that the it can be trained on the fly. Finally, we envision in situ testing for this incarnation of the robot.

Conclusion

We have detailed a simulation that implements a case-supported, principle-based behavior paradigm in the domain of eldercare including a subset of the perceptions and duties that will be required of a robot functioning in this domain. An example experiment that successfully determined the levels of duty satisfaction/violation for each action from perceptions was further described.

We have detailed a simulation of a robot, functioning in the domain of eldercare, whose behavior is completely determined by an ethical principle. Using a subset of the perceptions and duties that will be required of such a robot, we have demonstrated that this simulation selects ethically preferable actions in real time using a case-supported principle-based paradigm.

The ultimate goal of this project is to determine all the ethically relevant features and prima facie duties in the domain of eldercare, discover the principle that correctly balances them, and instantiate this principle in a full-featured eldercare robot. We believe that the effort required for this undertaking is worthwhile given that its fruit could serve as the basis for ensuring that the behavior of all eldercare robots that are created in the future will be ethically justifiable. Further, we believe that the methods used in this project can be employed in other domains as well, ensuring that robots that humans interact within these domains will behave ethically.

Acknowledgments

This material is based in part upon work supported by the NSF under Grant Numbers IIS-0500133, IIS-1151305 and IIS-1449155. We also wish to acknowledge the programming contributions of Patrick Tong, Robert Stolarz, Jameson Walsh, and Mat Rodrigue.

References

Abel, D., MacGlashan, J. and Littman, M. L., Reinforcement Learning as a Framework for Ethical Decision Making, AAAI-16 Workshop on AI, Ethics & Society, Technical Report WS-16-02, 2016.

Anderson, M. & Anderson, S. L., Machine Ethics: Creating an Ethical Intelligent Agent, AI Magazine, 28:4, Winter 2007.

Anderson, M. & Anderson, S. L., Robot Be Good, Scientific American 303.4 2010: 72-77.

Anderson, M. & Anderson, S. L., GenEth: A General Ethical Dilemma Analyzer, Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

Anderson, M. & Anderson, S. L., Toward Ensuring Ethical Behavior from Autonomous Systems: A Case-Supported Principle-Based Paradigm, Industrial Robot: An International Journal 2015 42:4, 324-331.

Arkin, R. C., Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture, Technical Report GIT-GVU-07-11, Georgia Institute of Technology.

Berenz, V., Tanaka, F., Suzuki, K., Herink, M. TDM: A software framework for elegant and rapid development of autonomous behaviors for humanoid robots. In Humanoids, pages: 179-186, IEEE, 2011

Berenz, V., Suzuki, K. Usability benchmarks of the Targets-Drives-Means robotic architecture In 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012), Osaka, Japan, November 29 - Dec. 1, 2012, pages: 514-519, 2012

Berenz, V., Suzuki, K. Targets-Drives-Means: A declarative approach to dynamic behavior specification with higher usability Robotics and Autonomous Systems, 62(4):545-555, 2014

Gruebler, A., Berenz, V., Suzuki, K. Emotionally Assisted Human-Robot Interaction Using a Wearable Device for Reading Facial Expressions, Advanced Robotics, 26(10):1143-1159, 2012

Kortenkamp, David and Simmons, Reid. Robotic systems architectures and programming. In Springer Handbook of Robotics, pages 187–206. Springer, 2008

Kuipers, B., Human-like morality and ethics for robots, AAAI-16 Workshop on AI, Ethics & Society, Technical Report WS-16-02, 2016.

Vanderelst, D. and Winfield, A., Architecture for Ethical Robots, https://arxiv.org/pdf/1609.02931.pdf, 2016