Ensuring Ethical Behavior from Autonomous Systems: A Case-Supported Principle-Based Paradigm

Michael Anderson
University of Hartford

Susan Leigh Anderson
University of Connecticut

Vincent Berenz
Max Planck Institute

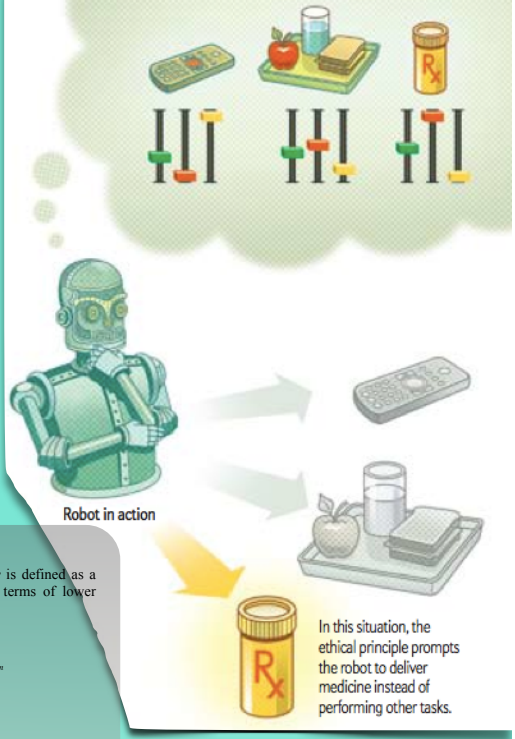
Research supported by NSF Grants IIS- 0500133, IIS-1151305, and IIS-1449155

Problem

A wide of variety of systems that interact with human beings are on the verge of being deployed in a number of domains (e.g. personal assistance, healthcare, driverless cars, search and rescue, etc.) and will be expected to navigate this ethically charged landscape responsibly.

Proposed Solution

Guide the behavior of such systems using explicitly represented ethical principles abstracted from a consensus of ethicists' judgements concerning particular cases.



Representation

From the perspective of ethics, actions can be characterized solely by the ***degrees*** of presence or absence of the ***ethically relevant features*** they involve and, indirectly, the *prima facie* ***duties*** they satisfy or violate.

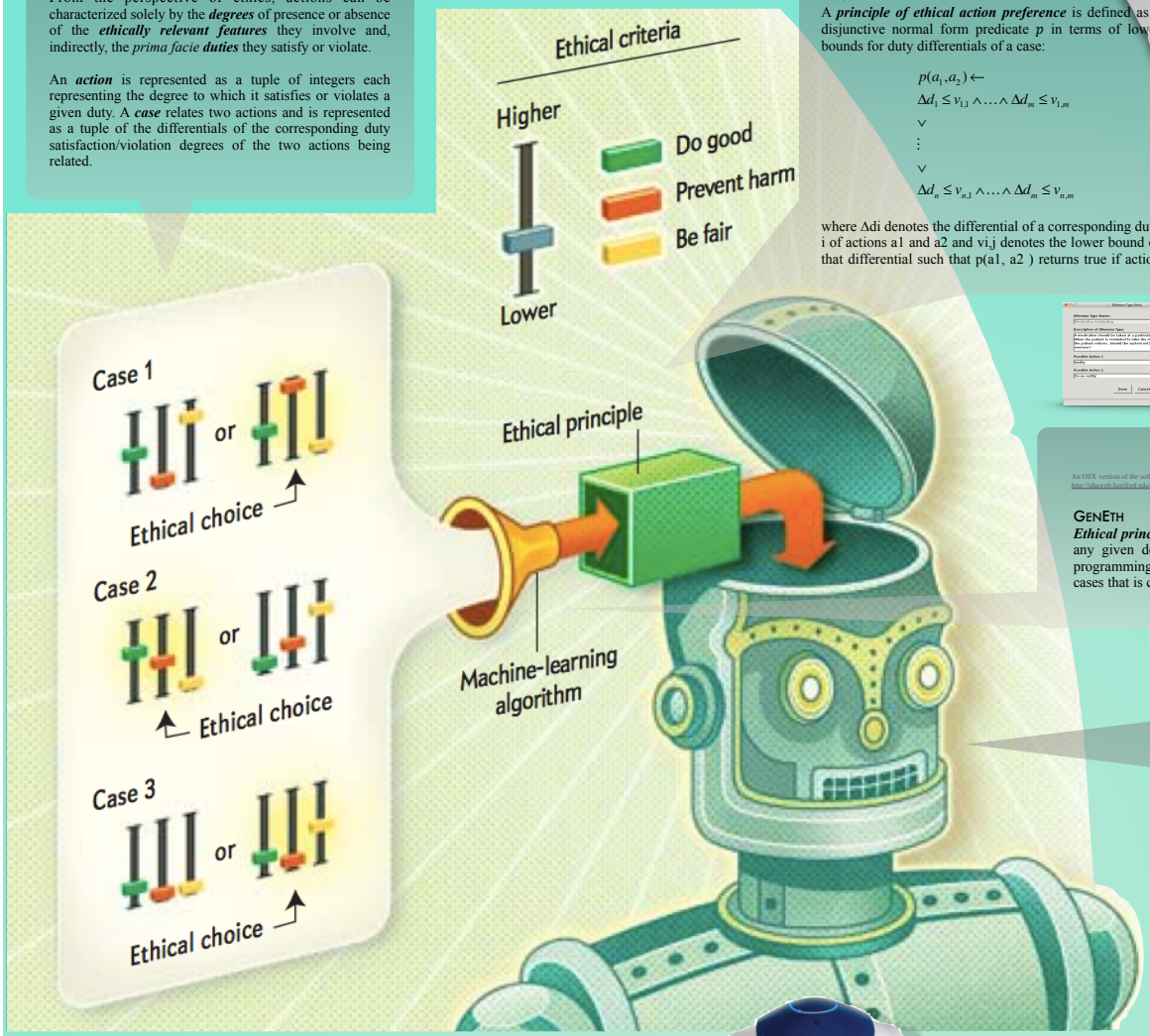
An ***action*** is represented as a tuple of integers each representing the degree to which it satisfies or violates a given duty. A ***case*** relates two actions and is represented as a tuple of the differentials of the corresponding duty satisfaction/violation degrees of the two actions being related.

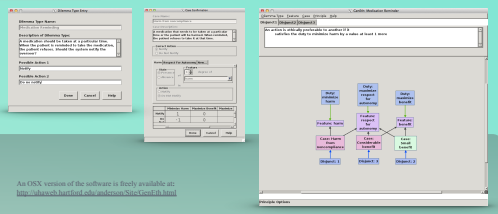
Principle

A ***principle of ethical action preference*** is defined as a disjunctive normal form predicate p in terms of lower bounds for duty differentials of a case:

$$p(a_1, a_2) \leftarrow$$
$$\Delta d_1 \leq v_{1,1} \wedge \ldots \wedge \Delta d_m \leq v_{1,m}$$
$$\vee$$
$$\vdots$$
$$\vee$$
$$\Delta d_n \leq v_{n,1} \wedge \ldots \wedge \Delta d_m \leq v_{n,m}$$

where Δdi denotes the differential of a corresponding duty i of actions a1 and a2 and vi,j denotes the lower bound of that differential such that p(a1, a2) returns true if action





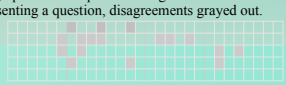
GenEth

Ethical principles are abstracted from specific cases of ethical dilemmas in any given domain, through a dialog with ethicists, using inductive logic programming (ILP) to infer a principle of ethical action preference from these cases that is complete and consistent in relation to them.

Evaluation

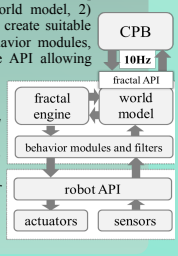
We have developed and administered an Ethical Turing Test that compares the ethically-preferable action specified by an ethicist in an ethical dilemma with that of a machine faced with the same dilemma. If a significant number of answers given by the machine match the answers given by the ethicist, then it has passed the test.

The test administered was comprised of 28 multiple-choice questions in four domains, one for each principle that was codified by the GenEth. These questions are drawn both from training (60%) and non-training cases (40%). It was administered to five ethicists, one of whom served as the ethicist on the project. ***Of the 140 questions, the ethicists agreed with the system's judgment on 123 of them or about 88% of the time.*** The graph below depicts this agreement with each cell representing a question, disagreements grayed out.



Fractal

Because autonomous robots are complex dynamic systems that must enforce stable control loops between sensors, estimated world model and action, integration of decision systems and high level behaviors into robots is a challenging task. This holds especially when human-robot interaction is one of the objectives, as the resulting robotic behavior has to look natural to any external observer. To deal with this complexity, ***we interfaced CPB with Fractal, our state of the art customizable robotic architecture that allows easy implementation of complex dynamic behaviors.*** It transparently: 1) implements the sensor filters required continuously to maintain an estimation of the world model, 2) adapts the layout of its program during runtime to create suitable data flow between decision, world model and behavior modules, and 3) provides its client software with a simple API allowing manipulation of a library of high level preemptive behaviors. Fractal is an extension of Targets-Drives-Means, a robotic architecture characterized by its high usability. ***Interfacing between CPB and Fractal allows the ethical decision procedure to run at a frequency of the order of 10 Hz, ensuring smooth execution of robotic behavior as well as a rapid runtime adaptation of the ethical behavior of the robot upon change in the situation.***



Implementation

Currently, we are using our general ethical dilemma analyzer (GenEth) to develop an ethical principle to guide the behavior of a Nao robot in the domain of eldercare. The robot's current set of possible actions includes charging, reminding a patient to take his/her medication, seeking tasks, engaging with patient, warning a non-compliant patient, and notifying an overseer. Sensory data such as battery level, motion detection, vocal responses, and visual imagery as well as overseer input regarding an eldercare patient are used to determine values for action duties pertinent to the domain. Currently these include maximize honor commitments, maximize readiness, minimize harm, maximize possible good, minimize non-interaction, maximize respect for autonomy, and minimize persistent immobility.

The robot's behavior at any given time is determined by sorting the actions by their ethical preference (represented by their duty values) and choosing the highest ranked one. As the learned principle returns true if the first of a pair of actions is ethically preferable to the second, it can be used as the comparison relation required by such sorting

Illustrations from Anderson and Anderson, "Robot Be Good", Scientific American, October 2010.