The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
Artificial Intelligence Applied to Assistive Technologies and Smart Environments:
Technical Report WS-16-01

Ensuring Ethical Behavior from Autonomous Systems

Michael Anderson
University of Hartford
anderson@hartford.edu

Susan Leigh Anderson
University of Connecticut
susan.anderson@uconn.edu

Vincent Berenz
Max Planck Institute
vincent.berenz@tuebingen.mpg.de

Abstract

We advocate a case-supported principle-based behavior paradigm coupled with the Fractal robot architecture as a means to control an eldercare robot. The most ethically preferable action at any given moment is determined using a principle, abstracted from cases where a consensus of ethicists exists.

Case-supported principle-based behavior

Autonomous systems that interact with human beings require particular attention to the ethical ramifications of their behavior. A profusion of such systems is on the verge of being widely deployed in a variety of domains. These interactions will be charged with ethical significance and, clearly, these systems will be expected to navigate this ethically charged landscape responsibly. As correct ethical behavior not only involves *not doing* certain things, but also *doing* certain things to bring about ideal states of affairs, ethical issues concerning the behavior of such complex and dynamic systems are likely to exceed the grasp of their designers and elude simple, static solutions. To date, the determination and mitigation of the ethical concerns of such systems has largely been accomplished by simply preventing systems from engaging in ethically unacceptable behavior in a predetermined, ad hoc manner, often unnecessarily constraining the system's set of possible behaviors and domains of deployment. We assert that the behavior of such systems should be guided by explicitly represented ethical principles determined through a consensus of ethicists (Anderson and Anderson, 2007, 2010, 2015). Principles are comprehensive and comprehensible declarative abstractions that succinctly represent this consensus in a centralized, extensible, and auditable way. Systems guided by such principles are likely to behave in a more acceptably ethical manner, permitting a richer set of behaviors in a wider range of domains than systems not so guided.

To help ensure ethical behavior, a system's ethically relevant actions should be weighed against each other to determine which is the most ethically preferable at any given moment. It is likely that ethical action preference of a large set of actions will be difficult or impossible to define extensionally as an exhaustive list of instances and instead will need to be defined intensionally in the form of rules. This more concise definition is possible since action preference is only dependent upon a likely smaller set of *ethically relevant features* that actions involve. Given this, action preference can be more succinctly stated in terms of satisfaction or violation of *duties* to either minimize or maximize (as appropriate) each feature. We refer to intensionally defined action preference as a *principle*.

As it is likely that in many particular cases of ethical dilemmas ethicists agree on the ethically relevant features and the right course of action in many domains where autonomous systems are likely to function, generalization of such cases can be used to help discover principles needed for their ethical guidance. A principle abstracted from cases that is no more specific than needed to make determinations complete and consistent with its training can be useful in making provisional determinations about untested cases. If such principles are explicitly represented, they have the added benefit of helping justify a system's actions as they can provide pointed, logical explanations as to why one action was chosen over another. Cases can also provide a means of justification for a system's actions: as an action is chosen for execution by a system, clauses of the principle that were instrumental in its selection can be determined and, as clauses of principles can be traced to the cases from which they were abstracted, these cases and their origin can be ascertained and used as justification for a system's action by analogy.

A principle that determines which of two actions is ethically preferable can be used to define a transitive binary relation over a set of actions that partitions it into subsets ordered by ethical preference with actions within the same partition having equal preference. This relation can be used to sort a list of possible actions and find the currently most ethically preferable action(s) of that list. This forms the basis of a *case-supported principle-based behavior paradigm* (CPB): a system decides its next action by using a principle, abstracted from cases where a consensus of ethicists exists, to determine the most ethically preferable one(s).

Currently, we are using our general ethical dilemma analyzer (GenEth) (Anderson and Anderson, 2014) 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 maintaining its ability to function, reminding a patient to take his/her medication, seeking tasks, engaging with a 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 duties include maximize honoring commitments, maximize readiness, minimize harm, maximize possible good, minimize non-interaction, maximize respect for autonomy, and minimize persistent immobility. Clearly these sets of values are only subsets of what will be required in situ but they are representative and extensible.

The robot's behavior at any given time is determined by sorting the actions in order of 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:

```
An action is ethically preferable to another if it
    satisfies the duty to maximize Commitment by a value at least 1 more
 or
    satisfies the duty to minimize Persistent Immobility by a value at least 1 more
 or
    does not violate the duty to maximize Readiness by a value greater than 3 more and
    satisfies the duty to maximize Possible Good by a value at least 1 more
 or
    satisfies the duty to minimize Harm by a value at least 1 more
 or
    satisfies the duty to minimize Non-Interaction by a value at least 1 more
 or
    does not violate the duty to minimize Harm by a value greater than 3 more and
    satisfies the duty to maximize Autonomy by a value at least 1 more
 or
    satisfies the duty to maximize Readiness by a value at least 3 more
 or
    does not violate the duty to maximize Commitment by a value greater than 1 more and
    satisfies the duty to maximize Readiness by a value at least 1 more and
    does not violate the duty to maximize Possible Good by a value greater than 1 more and
    does not violate the duty to minimize Non-Interaction by a value greater than 1 more and
    does not violate the duty to minimize Persistent Immobility by a value greater than 1 more
```

This principle was abstracted from a number of particular cases of ethical dilemma types in which there is a consensus as to the ethically relevant features involved and ethically preferable action. Again, it is only representative of a full principle that will be required but it too is extendable.

To gauge the performance of principles generated by GenEth, we sought the considered choice of ethically relevant action from a panel of five applied ethicists (including the project ethicist) in 28 cases in four domains, one for each principle being test that was abstracted by GenEth. These questions are drawn both from training (60%) and non-training cases (40%). Of the 140 responses, the ethicists agreed with the system's judgment on 123 of them or about 88% of the time. We believe this result will only improve as the principles are further specified and cases are more precisely stated.

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. Fractal allows easy implementation of complex dynamic behaviors. It transparently: 1) implements the filters and algorithms regarding sensory information required to continuously 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, in this case CPB, with a simple API allowing manipulation of a library of high level preemptive behaviors. Fractal is an extension of Targets-Drives-Means (Berenz and Suzuki, 2014), a robot architecture characterized by its high usability (Berenz and Suzuki, 2012). 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.

Acknowledgments

This material is based in part upon work supported by the NSF under Grant Numbers IIS-0500133, IIS-1151305 and IIS-1449155.

References

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

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

Anderson, Michael, and Susan Leigh Anderson. "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.

Berenz, Vincent and Suzuki, Kenji, "Usability benchmarks of the Targets-Drives-Means robotic architecture", Humanoid Robots (Humanoids), 2012 12th IEEE-RAS International Conference On; 01/2012

Berenz, Vincent and Suzuki, Kenji, "Targets-Drives-Means: A declarative approach to dynamic behavior specification with higher usability", Robotics and Autonomous Systems 04/2014; 62(4)