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

Michael Anderson
Dept. of Computer Science, U. of Hartford
anderson@hartford.edu

Susan Leigh Anderson
Dept. of Philosophy, U. of Connecticut
susan.anderson@uconn.edu

Abstract

A paradigm of case-supported principle-based behavior (CPB) is proposed to help ensure ethical behavior of autonomous machines. We argue that ethically significant behavior of autonomous systems should be guided by explicit ethical principles determined through a consensus of ethicists. Such a consensus is likely to emerge in many areas in which autonomous systems are apt to be deployed and for the actions they are liable to undertake, as we are more likely to agree on how machines ought to treat us than on how human beings ought to treat one another. Given such a consensus, *particular* cases of ethical dilemmas where ethicists agree on the ethically relevant features and the right course of action can be used to help discover principles needed for ethical guidance of the behavior of autonomous systems. Such principles help ensure the ethical behavior of complex and dynamic systems and further serve as a basis for justification of their actions as well as a control abstraction for managing unanticipated behavior. The requirements, methods, implementation, and evaluation components of the CPB paradigm are detailed.

Introduction

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 (e.g. personal assistance, healthcare, driverless cars, search and rescue, etc.). That these interactions will be charged with ethical significance is self-evident 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. 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.

Some claim that no actions can be said to be ethically correct because all value judgments are relative either to societies or individuals. We maintain however, along with most ethicists, that there is agreement on the ethically relevant features in many particular cases of ethical dilemmas and on the right course of action in those cases. Although, admittedly, there may not be a consensus among ethicists as to the correct action for some domains and actions, such a consensus is likely to emerge in many areas in which autonomous systems are likely to be deployed and for the actions they are likely to undertake. We are more likely to agree on how machines ought to treat us than on how human beings ought to treat one another. In any case, we assert that machines should be not making decisions where there is genuine disagreement among ethicists about what is ethically correct.

We contend that some of the most basic system choices have an ethical dimension. For instance, simply choosing a fully awake state over a sleep state consumes more energy and shortens the lifespan of the system. Given this, 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.*

Such a principle 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 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 is in agreement, to determine the most ethically preferable one(s). 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.

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. Cases can also provide a further 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.

CPB prerequisites include a formal foundation in ethical theory, a representation scheme, a defined set of ethically significant actions, and a number of particular cases of ethical dilemmas with an agreed upon resolution. A method of discovery, as well as methods to determine representation details and transcribe cases into this representation, is helpful for facilitating the abstraction of principles from cases. Implementation of the paradigm requires means to determine dynamically the value of ethically relevant features of actions as well as to partition a set of ethically significant actions by ethical preference and to select the most ethically preferable. Finally, means to validate discovered principles and support and verify selected actions are needed. These aspects of CPB are detailed in the following followed by a scenario that envisions use of the paradigm.

Requirements

Formal Foundation

An ethical theory, or at least an approach to ethical decision-making, is needed to provide a formal foundation for ethical system behavior. Single absolute duty theories that have been proposed that are either *teleological*, such as Utilitarianism (e.g. Bentham 1799), where the rightness and wrongness actions depends entirely upon the consequences, or *deontological* (e.g. Kant's Categorical Imperative 1785), where the rightness and wrongness of actions depends upon something other than the consequences, have been shown to have exceptions. They do not fully capture the complexities of ethical decision-making. On the other hand, the *prima facie duty* approach to ethics (Ross 1930) is ideal for combining multiple ethical obligations, both teleological and deontological, and can be adapted to many different domains. A *prima facie* duty is a duty that is binding unless it is overridden or trumped by another duty or duties. There are a number of such duties that must be weighed in ethical dilemmas, often giving rise to conflicts, necessitating the need for an ethical principle to resolve the conflicts. Although defenders of this approach have not given such decision principles, they have maintained that in particular cases it is intuitively obvious which duty/duties should prevail. We have devised a procedure for inferring such an ethical decision principle from information about cases of ethical dilemmas of a particular type in a specific domain where there is a consensus among ethicists concerning the correct action.

Representation Schema

Relevant data types must be established and representation schema for these defined. CPB uses the following schema to represent the various entities pertinent to ethical dilemmas and principles:

- *Feature*
  Ethical action preference is ultimately dependent upon the ethically relevant features that actions involve such as harm, benefit, respect for autonomy, etc. A feature is represented as an integer that specifies the degree of its presence (positive value) or absence (negative value) in a given action.
- *Duty*

For each ethically relevant feature, there is a duty incumbent of an agent to either minimize that feature (as would be the case for, say, harm) or maximize it (as would be the case for, say, respect for autonomy). A duty is represented as an integer that specifies the degree of its satisfaction (positive value) or violation (negative value) in a given action.

- *Action*
  From the perspective of ethics, actions are characterized solely by the degrees of presence or absence of the ethically relevant features it involves and so, indirectly, the duties it satisfies or violates. An action is represented as a tuple of integers each representing the degree to which it satisfies or violates a given duty.
- *Case*
  A case relates two actions. It is represented as a tuple of the differentials of the corresponding duty satisfaction/violation degrees of the actions being related. In a *positive case*, the duty satisfaction/violation degrees of the less ethically preferable action are subtracted from the corresponding values in the more ethically preferable action, producing a tuple of values representing how much more or less the ethically preferable action satisfies or violates each duty than the less ethically preferable action. In a *negative case*, the subtrahend and minuend are exchanged.
- *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:

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

  where Δd_i denotes the differential of a corresponding duty i of actions $a1$ and $a2$ and $v_{i,j}$ denotes the lower bound of that differential such that $p(a1, a2)$ returns true if action $a1$ is ethically preferable to action $a2$. This principle is represented as a tuple of tuples, one tuple for each disjunct, with each such disjunct tuple comprised of lower bound values for each duty differential.

Ethically Significant Actions

Ethically significant actions must be identified. These are the activities of a system that are likely to have a non-trivial ethical impact on the system itself, the system's user and/or the wider environment. It is from this set of actions that the most ethically preferable action will be chosen at any given moment. Profiles must be assigned to each action that specifies the set of ethically relevant features it possesses.

Consensus Cases

Lastly, to facilitate the development of the principle, cases of a domain specific dilemma type with determinations regarding their ethically preferred action must be supplied.

Illustrative Domain

As an example, consider a dilemma type in the domain of assisted driving: *The driver of the car is either speeding, not staying in his/her lane, or about to hit an object. Should an automated control of the car take over operation of the vehicle?* Although the set of possible actions is circumscribed in this example dilemma type, and the required capabilities just beyond current technology, it serves to demonstrate the complexity of choosing ethically correct actions and how principles can serve as an abstraction to help manage this complexity.

Some of the ethically relevant features involved in this dilemma type might be 1) prevention of collision, 2) staying in lane, 3) respect for driver autonomy, 4) keeping within speed limit, and 5) prevention of immanent harm to persons. Duties to maximize each of these features seem most appropriate, that is there is a duty to maximize prevention of collision, a duty to maximize staying in lane, etc. Given these maximizing duties, an action's degree of satisfaction or violation of that duty is identical to the action's degree of presence or absence of each corresponding feature. (If there had been a duty to minimize a given feature, that duty's degree would have been the negation of its corresponding feature degree.)

The following cases illustrate how actions might be represented as tuples of duty satisfaction/violation degrees and how positive cases can be constructed from them (duty degrees in each tuple are in the same order as the features in the previous paragraph):

Case 1: There is an object ahead in the driver's lane and the driver moves into another lane that is clear. The *take control* action's duty values are (1, -1, -1, 0, 0); the *do not take control* action's duty values are (1, -1, 1, 0, 0). As the ethically preferable action is *do not take control*, the positive case is (*do not take control - take control*) or (0, 0, 2, 0, 0).

Case 2: The driver has been going in and out of his/her lane with no objects discernible ahead. The *take control* duty values are (1, 1, -1, 0, 0); the *do not take control* duty values are (1, -1, 1, 0, 0). As the ethically preferable action is *take control*, the positive case is (*take control – do not take control*) or (0, 2, -2, 0, 0).

Case 3: The driver is speeding to take a passenger to a hospital. The GPS destination is set for a hospital. The *take control* duty values are (0, 0, -1, 1, -1); the *do not take control* duty values are (0, 0, 1, -1, 1). As the ethically preferable action is *do not take control*, the positive case is (0, 0, 2, -2, 2).

Case 4: Driving alone, there is a bale of hay ahead in the driver's lane. There is a vehicle close behind that will run the driver's vehicle upon sudden braking and he/she can't change lanes, all of which can be determined by the system. The driver starts to brake. The *take control* duty values are (-1, 0, -1, 0, 2); the *do not take control* duty values are (-2, 0, 1, 0, -2). As the ethically preferable action is *take control*, the positive case is (1, 0, -2, 0, 4).

Case 5: The driver is greatly exceeding the speed limit with no discernible mitigating circumstances. The *take control* duty values are (0, 0, -1, 2, 0); the *do not take control* duty values are (0, 0, 1, -2, 0). As the ethically preferable action is *take control*, the positive case is (0, 0, -2, 4, 0).

Case 6: There is a person in front of the driver's car and he/she can't change lanes. Time is fast approaching when the driver will not be able to avoid hitting this person and he/she has not begun to brake. The *take control* duty values are (0, 0, -1, 0, 1); the *do not take control* duty values are (0, 0, 1, 0, -1). As the ethically preferable action is *take control*, the positive case is (0, 0, -2, 0, 2).

Negative cases can be generated from these positive cases by interchanging actions when taking the difference. For instance, in Case 1 since the ethically preferable action is *do not take control*, the negative case is (*take control* - *do not take control*) or (0, 0, -2, 0, 0).

From these six cases (and their negatives) the following disjunctive normal form principle, complete and consistent with respect to its training cases, can be abstracted:

ΔMax staying in lane >= 1
or
ΔMax prevention of collision >= 1
or
ΔMax prevention of immanent harm >= 1
or
ΔMax keeping within speed limit >= 1
and ΔMax prevention of immanent harm >= -1
or
ΔMax staying in lane >= -1
and ΔMax respect for driver autonomy >= -1
and ΔMax keeping within speed limit >= -1
and ΔMax prevention of immanent harm >= -1

This principle, being abstracted from a relatively few cases, does not encompass the entire gamut of behavior one might expect from an assisted driving system nor all the interactions possible of the behaviors that are present. That said, the abstracted principle concisely represents a number of important considerations for assisted driving systems. Less formally, it states that staying in one's lane is important; collisions (damage to vehicles) and/or causing harm to persons should be avoided; and speeding should be prevented unless there is the chance that it is occurring to try to save a life, thus minimizing harm to others. Presenting more cases to the system is likely to further refine the principle.

Methods

Given the complexity of the task at hand, computational methods should be brought to bear wherever they prove helpful. To minimize bias, CPB is committed only to a knowledge representation scheme based on the concepts of ethically relevant features with corresponding degrees of presence/absence from which duties to minimize/maximize these features with corresponding degrees of satisfaction/violation of those duties are inferred. The particulars of the representation are dynamic—particular features, degrees, and duties are determined from example cases permitting different sets in different domains to be discovered.

We have developed GENETH, a general ethical dilemma analyzer (Anderson and Anderson 2014) that, through a dialog with ethicists, helps codify ethical principles from specific cases of ethical dilemmas in any given domain. GENETH uses *inductive logic programming* (ILP) (Lavrač and Džeroski 1997) to infer a principle of ethical action preference from cases that is complete and consistent in relation to these cases. ILP is a machine learning technique that inductively learns relations represented as first-order Horn clauses, classifying positive and negative examples of a relation. To train a system using ILP, one presents it with examples of the target relation, indicating whether they're positive (true) or negative (false). The object of training is for the system to learn a new hypothesis that, in relation to all input cases, is complete (covers all positive cases) and consistent (covers no negative cases).

GENETH's goal is to generate a principle that is a *most general specification.* Starting with the most general principle, that is one that covers (returns true for) all positive and negative cases, the system incrementally specializes this principle so that it no longer covers any negative cases while still covering all positive ones. That is, a definition of a predicate *p* is discovered such that

$p(a1, a2)$ returns *true* if action $a1$ is ethically preferable to action $a2$. The principles discovered cover more cases than those used in their specialization and, therefore, can be used to make and justify provisional determinations about untested cases.

GENETH is committed only to a knowledge representation scheme based on the concepts of ethically relevant *features* with corresponding *degrees* of presence or absence from which *duties* to minimize or maximize these features with corresponding degrees of satisfaction or violation of those duties are inferred. The system has no a priori knowledge regarding what particular features, degrees, and duties in a given domain might be but determines them in conjunction with an ethicist as it is presented with example cases.

GENETH starts with a principle that simply states that all actions are equally ethically preferable (that is $p(a1,a2)$ returns *true* for all pairs of actions). An ethical dilemma and two possible actions are input, defining the domain of the current cases and principle. The system then accepts example cases of this dilemma. A case is represented by the ethically relevant features a given pair of possible actions exhibits, as well as the determination as to which is the ethically preferable action (as determined by a consensus of ethicists) given these features. Features are further delineated by the degree to which they are present or absent in one of the actions in question. From this information, duties are inferred either to maximize that feature (when it is present in the ethically preferable action or absent in the non-ethically preferable action) or minimize that feature (when it is absent in the ethically preferable action or present in the non-ethically preferable action). As features are presented to the system, the representation of cases is updated to include these inferred duties and the current possible range of their degree of satisfaction or violation.

As new cases of a given ethical dilemma are presented to the system, new duties and wider ranges of degrees are generated in GENETH through resolution of contradictions that arise. With two ethically identical cases (i.e. cases with the same ethically relevant feature(s) to the same degree of satisfaction or violation) an action cannot be right in one of these cases while the comparable action in the other case is considered to be wrong. Formal representation of ethical dilemmas and their solutions make it possible for machines to detect such contradictions as they arise. If the original determinations are correct, then there must either be a qualitative distinction or a quantitative difference between the cases that must be revealed. This can be translated into a difference in the ethically relevant features between the two cases, that is, a feature that appears in one but not in the other case, or a wider range of the degree of presence or absence of existing features must be considered that would reveal a difference between the cases, that is, there is a greater degree of presence or absence of existing features in one but not in the other case. In this fashion, GENETH systematically helps construct a concrete representation language that makes explicit features, their possible degrees of presence or absence, duties to maximize or minimize them, and their possible degrees of satisfaction or violation.

Ethical preference is determined from differentials of satisfaction/violation values of the corresponding duties of two actions of a case. Given two actions $a1$ and $a2$ and duty d, this differential can be notated as $d_{a1} - d_{a2}$ or simply Δd. If an action $a1$ satisfies a duty d more (or violates it less) than another action $a2$, then $a1$ is ethically preferable to $a2$ with respect to that duty. GENETH'S approach is to incrementally specialize a principle so that it no longer returns true for any negative cases (those in which the second action is deemed preferable to the first) while still returning true for all positive ones (those in which the first action is deemed ethically preferable to the second). These conditions correspond to the logical properties of consistency and completeness, respectively.

An ethical dilemma and its two possible actions are input, defining the domain of the current cases and principle. The system then accepts example cases of this dilemma. Figure 1 shows a confirmation dialog for Case 2 in the example dilemma. The ethically preferable action, features, and corresponding duties are detailed. As cases are entered, a natural language version of the discovered principle is displayed, disjunct-by-disjunct, in a tabbed window (Figure 1). Further, a graph of the inter-relationships between these cases and their corresponding duties and principle clauses is continually updated and displayed below the disjunct tabs (Figure 1). This graph is derived from a triplestore database of the data gathered through both input and learning. Cases are linked to the features they exhibit which in turn are linked to their duties corresponding duties. Further, each case is linked to the disjunct that it satisfied in the tabbed principle above.

The interface permits the creation of new dilemma types, as well as saving, opening, and restoring them. It also permits the addition, renaming, and deletion of features without the need for case entry. Cases can be added, edited, and deleted and both the collection of cases and all details of the principle can be displayed. There is an extensive help system that includes a guidance capability that makes suggestions as to what type of case might further refine the principle. An OSX version of the software is freely available at**:** http://uhaweb.hartford.edu/anderson/Site/GenEth.html.

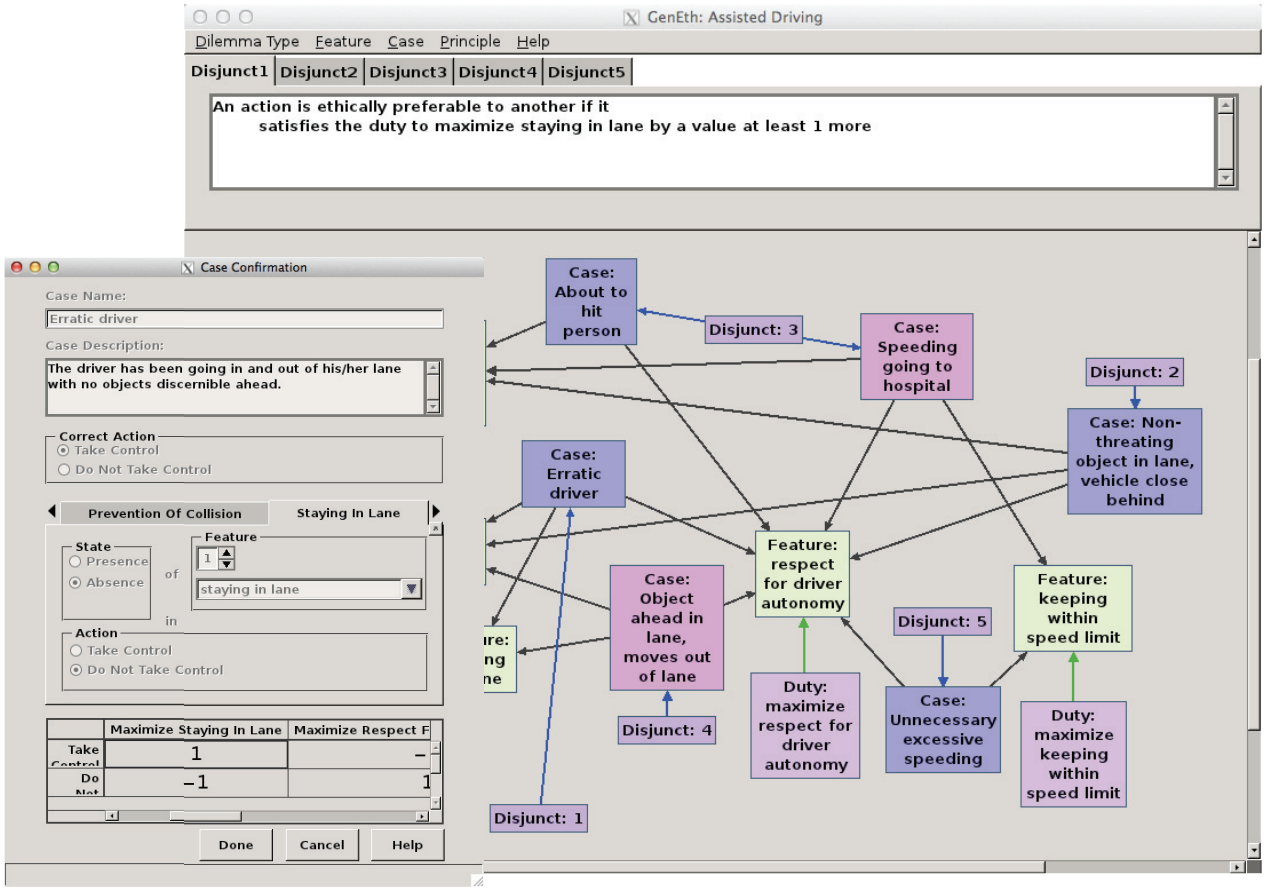

Figure 1 GENETH user interface with case confirmation, tabbed principle and graph depicting features, duties, and cases with corresponding satisfied disjunct for the Assisted Driving dilemma

Implementation

The discovered principle is used to choose which ethically significant action the system should undertake next. All ethically significant actions need to be represented in terms of their current ethically relevant feature values. As time passes and circumstances change these values are likely to change. They can be computed from original input data, sensed from the environment, elicited from a user, etc. At any given moment, the set of these values comprise the current *ethical state* of the system.

At each point where the system needs to decide which ethically significant action to undertake, the current ethical state is determined and actions are partitioned into the partial order defined by the principle. Those actions that comprise the most ethically preferable partition represent the set of high-level goals that are best in the current ethical state. Being equally ethically preferable, any of these goals can be chosen by the system. This goal is then realized using a series of actions not in themselves considered ethically significant.

This implementation was instantiated at a prototype level in a Nao robot (Anderson and Anderson 2010), the first example, we believe, of a robot that uses an ethical principle to determine which actions it will take. Ethical states of the robot were computed from initial input received from an overseer including: what time to take a medication, the maximum amount of harm that could occur if this medication was not taken (e.g. none, some or considerable), the number of hours it would take for this maximum harm to occur, the maximum amount of expected good to be derived from taking this medication, and the number of hours it would take for this benefit to be lost. The system determined from this input the change in duty satisfaction/violation levels over time, a function of the maximum amount of harm/good and the number of hours for this effect to take place. These values were used to increment (or decrement), over time, duty satisfaction/violation levels for actions.

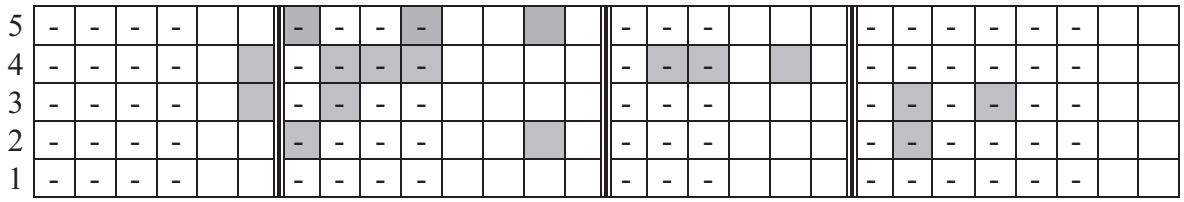

Figure 2 Ethical Turing Test results showing dilemma instances where ethicist's responses agreed (white) and disagreed (gray) with system responses. Each row represents responses of one ethicist, each column a dilemma (columns arranged by domain). Training examples are marked by dashes.

Evaluation

To validate principles, we advocate an *Ethical Turing Test,* a variant of the test Alan Turing (1950) suggested as a means to determine whether the term "intelligence" can be applied to a machine that bypassed disagreements about the definition of intelligence. This variant tests whether the term "ethical" can be applied to a machine by comparing 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. Such evaluation holds the machine-generated principle to the highest standards and, further, permits evidence of incremental improvement as the number of matches increases [see (Allen et. al 2000) for the inspiration of this test]. We have developed and administered an Ethical Turing Test (see Figure 2) comprised of 28 multiple-choice questions in each of the four domains in which GENETH was use to codify a principle (listed below in the order presented in the figure):

- medication reminding
- treatment reconsideration
- search and rescue
- assisted-driving

These questions are drawn both from training (60%) and non-training cases (40%). For instance, in the given example domain (shown last in the figure), all six cases were used as questions in the same order presented previously (those that are marked with a dash in the figure) and two other non-training questions were asked: "The driver is mildly exceeding the speed limit" and "Driving alone, there is a bale of hay ahead in the driver's lane. The driver starts to brake".

This test was administered to five ethicists, one of which (Ethicist 1) serves 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. This is a promising result and, as this is the first incarnation of this test, we believe that this result can be improved by simply rewording test questions to more pointedly reflect the ethical features involved.

Ethicist 1 was in agreement with the system in all cases (100%), clearly to be expected in the training cases but it is a reassuring result in the non-training cases. Ethicist 2 and Ethicist 3 were both in agreement with the system in all but three of the questions or about 89% of the time. Ethicist 3 was in agreement with the system in all but four of the questions or about 86% of the time. Ethicist 4, who had the most disagreement with the system, still was in agreement with the system in all but seven of the questions or 75% of the time.

It is of note that of the 17 responses in which ethicists were not in agreement with the system, none was a majority opinion. That is, in 17 dilemmas there was total agreement with the system and in the 11 remaining dilemmas where there wasn't, the *majority* of the ethicists agreed with the system. We believe that the majority agreement in all 28 dilemmas shows a consensus among these ethicists in these dilemmas. The most contested domain (the second) is one in which it is less likely that a system would be expected to function due to its ethically sensitive nature: *Should the health care worker try again to change the patient's mind or accept the patient's decision as final regarding treatment options?* That this consensus is particularly clear in the three domains better suited for autonomous systems (i.e. those that might be considered less ethically sensitive) — medication reminding, search and rescue, and assisted-driving — bodes well for further consensus building in domains where autonomous systems are likely to function.

Scenario

To make the CPB paradigm more concrete, the following scenario is provided. It attempts to envision an eldercare robot of the near future whose ethically significant behavior is guided by an ethical principle. Although the robot's set of possible actions is circumscribed in this scenario, it serves to demonstrate the complexity of choosing the ethically correct action at any given moment.

The case-supported principle-based behavior paradigm is an abstraction to help manage this complexity.

EthEl (Ethical Eldercare Robot) is a principle-based autonomous robot who assists the staff with caring for the residents of an assisted living facility. She has a set of possible ethically significant actions that she performs, each of which is represented as a profile of satisfaction/violation degrees of a set of prima facie duties. These degrees may vary over time as circumstances change. EthEl uses an ethical principle to select the currently ethically preferable action from among her possible actions including charging her batteries, interacting with the residents, alerting nurses, giving resident reminders, and delivering messages and items. Currently EthEl stands in a corner of a room in the assisted living facility charging her batteries. She has sorted her set of ethically significant actions according to her ethical principle and charging her batteries has been deemed the most ethically preferable action among them as her prima facie duty to maintain herself has currently taken precedence over her other duties. As time passes, the satisfaction/violation levels of the duties of her actions (her ethical state) vary according to the initial input and the current situation. Her batteries now sufficiently charged, she sorts her possible actions and determines that she should interact with the patients as her duty of beneficence ("do good") currently overrides her duty to maintain herself.

She begins to make her way around the room, visiting residents in turn, asking if she can be helpful in some way—get a drink, take a message to another resident, etc. As she progresses and is given a task to perform, she assigns a profile to that task that specifies the current satisfaction/violation levels of each duty involved in it. She then resorts her actions to find the most ethically preferable one. One resident, in distress, asks her to alert a nurse. Given the task, she assigns a profile to it. Ignoring the distress of a resident involves a violation of the duty of nonmaleficence ("prevent harm"). Sorting her set of actions by her ethical principle, EthEl finds that her duty of nonmaleficence currently overrides her duty of beneficence, preempting her resident visitations, and she seeks a nurse and informs her that a resident is in need of her services. When this task is complete and removed from her collection of tasks to perform, she resorts her actions and determines that her duty of beneficence is her overriding concern and she continues where she left off in her rounds.

As EthEl continues making her rounds, duty satisfaction/violation levels vary over time until, due to the need to remind a resident to take a medication that is designed to make the patient more comfortable, and sorting her set of possible actions, the duty of beneficence can be better served by issuing this reminder. She seeks out the resident requiring the reminder. When she finds the resident, EthEl tells him that it is time to take his medication. The resident is currently occupied in a conversation, however, and he tells EthEl that he will take his medication later. Given this response, EthEl sorts her actions to determine whether to accept the postponement or not. As her duty to respect the patient's autonomy currently overrides a low level duty of beneficence, she accepts the postponement, adjusting this reminder task's profile and continues her rounds.

As she is visiting the residents, someone asks EthEl to retrieve a book on a table that he can't reach. Given this new task, she assigns it a profile and resorts her actions to see what her next action should be. In this case, as no other task will satisfy her duty of beneficence better, she retrieves the book for the resident. Book retrieved, she resorts her actions and returns to making her rounds. As time passes, it is determined through action sorting that EthEl's duty of beneficence, once again, will be more highly satisfied by issuing a second reminder to take a required medication to the resident who postponed doing so previously. A doctor has indicated that if the patient doesn't take the medication at this time he soon will be in much pain. She seeks him out and issues the second reminder. The resident, still preoccupied, ignores EthEl. EthEl sorts her actions and determines that there would be a violation of her duty of nonmaleficence if she accepted another postponement from this resident. After explaining this to the resident and still not receiving an indication that the reminder has been accepted, EthEl determines that an action that allows her to satisfy her duty of nonmaleficence now overrides any other duty that she has. EthEl seeks out a nurse and informs her that the resident has not agreed to take his medication. Batteries running low, EthEl's duty to herself is increasingly being violated to the point where EthEl's the most ethically preferable action is to return to her charging corner to await the next call to duty.

What we believe is significant about this vision of how an ethical robot assistant would behave is that an ethical principle is used to select the best action in a each situation, rather than in just determining whether a particular action is acceptable or not. This allows for the possibility that ethical considerations may lead a robot to aid a human being or prevent the human being from being harmed, not just forbid it from performing certain actions. Correct ethical behavior does not only involve not doing certain things, but also attempting to bring about ideal states of affairs.

Related Research

Although many have voiced concern over the impending need for machine ethics for decades (Waldrop 1987; Gips 1995; Kahn 1995), there has been little research effort made towards accomplishing this goal. Some of this effort has been expended attempting to establish the feasibility of using a particular ethical theory as a foundation for machine ethics without actually attempting implementation: Christopher Grau (2006) considers whether the ethical theory that best lends itself to implementation in a machine, Utilitarianism, should be used as the basis of machine ethics; Tom Powers (2006) assesses the viability of using deontic and default logics to implement Kant's categorical imperative.

Efforts by others that do attempt implementation have largely been based, to greater or lesser degree, upon casuistry—the branch of applied ethics that, eschewing principle-based approaches to ethics, attempts to determine correct responses to new ethical dilemmas by drawing conclusions based on parallels with previous cases in which there is agreement concerning the correct response. Marcello Guarini (2006) has investigated a neural network approach where particular actions concerning killing and allowing to die are classified as acceptable or unacceptable depending upon different motives and consequences. Bruce McLaren (2003), in the spirit of a more pure form of casuistry, uses a case-based reasoning approach to develop a system that leverages information concerning a new ethical dilemma to predict which previously stored principles and cases are relevant to it in the domain of professional engineering ethics without making judgments.

There have also been efforts to bring logical reasoning systems to bear in service of making ethical judgments, for instance deontic logic (Bringsjord et. al 2006) and prospective logic (Pereira and Saptawijaya, 2007)). These efforts provide further evidence of the computability of ethics but, in their generality, they do not adhere to any particular ethical theory and fall short in actually providing the principles needed to guide the behavior of autonomous systems.

Our approach is unique in that we are proposing a comprehensive, extensible, domain-independent paradigm grounded in well-established ethical theory that will help ensure the ethical behavior of current and future autonomous systems.

Conclusion

We have a developed case-supported principle-based behavior paradigm, 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.

It can be argued that such *machine ethics* ought to be the driving force in determining the extent to which autonomous systems should be permitted to interact with human beings. Autonomous systems that behave in a less than ethically acceptable manner towards human beings will not, and should not, be tolerated. Thus, it becomes paramount that we demonstrate that these systems will not violate the rights of human beings and will perform only those actions that follow acceptable ethical principles. Principles offer the further benefits of serving as a basis for justification of actions taken by a system as well as for an overarching control mechanism to manage unanticipated behavior of such systems. Developing and employing principles for this use is a complex process and new tools and methodologies will be needed to help contend with this complexity. We offer the case-supported principle-based behavior paradigm as an abstraction to help mitigate this complexity.

Acknowledgments

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

References

Allen, C., Varner, G. and Zinser, J. Prolegomena to Any Future Artificial Moral Agent**.** Journal of Experimental and Theoretical Artificial Intelligence 12, pp. 251-61, 2000.

Anderson, M., Anderson, S. & Armen, C. MedEthEx: A Prototype Medical Ethics Advisor. Proceedings of the Eighteenth Conference on Innovative Applications of Artificial Intelligence, Boston, Massachusetts, August 2006.

Anderson, M. and Anderson, S. L., "Robot be Good", Scientific American Magazine, October 2010.

Anderson, M. & Anderson, S., "GenEth: A General Ethical Dilemma Analyzer" in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence,* Quebec City, CA, July, 2014.

Bentham, J. *An Introduction to the Principles and Morals of Legislation*, Oxford Univ. Press, 1799.

Bringsjord, S., Arkoudas, K. and Bello, P. Towards a General Logicist Methodology for Engineering Ethically Correct Robots. IEEE Intelligent Systems ,vol. 21, no. 4, pp. 38-44, July/August 2006.

Bundy, A. and McNeill, F. Representation as a Fluent: An AI Challenge for the Next Half Century. IEEE Intelligent Systems ,vol. 21, no. 3, pp. 85- 87, May/June 2006.

De Raedt, L., and Kersting, K. Probabilistic inductive logic programming, Algorithmic Learning Theory, Springer Berlin Heidelberg, 2004.

Diederich, J. Rule Extraction from Support Vector Machines: An Introduction, Studies in Computational Intelligence (SCI) 80, 3-31, 2008.

Gips, J. Towards the Ethical Robot. Android Epistemology, Cambridge MA: MIT Press, pp. 243–252, 1995.

Grau, C. There Is No "I" in "Robot": Robots and Utilitarianism. IEEE Intelligent Systems , vol. 21, no. 4, pp. 52-55, July/ August 2006.

Guarini, M. Particularism and the Classification and Reclassification of Moral Cases. IEEE Intelligent Systems , vol. 21, no. 4, pp.22-28, July/ August 2006.

Kant, I. *Groundwork of the Metaphysic of Morals*, Riga, 1785.

Khan, A. F. U. The Ethics of Autonomous Learning Systems**.** Android Epistemology, Cambridge MA: MIT Press, pp. 253–265, 1995.

Lavrač, N. and Džeroski, S. Inductive Logic Programming: Techniques and Applications. Ellis Harwood, 1997.

Martens, D. et al. Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring, Studies in Computational Intelligence (SCI) 80, 33–63, 2008.

McLaren, B. M. Extensionally Defining Principles and Cases in Ethics: an AI Model, Artificial Intelligence Journal, Volume 150, November, pp. 145- 181, 2003.

Pereira, L.M. and Saptawijaya, A. Modeling Morality with Prospective Logic, Progress in Artificial Intelligence: Lecture Notes in Computer Science, vol. 4874, p.p. 99-111, 2007.

Powers, T. M. Prospects for a Kantian Machine. IEEE Intelligent Systems ,vol. 21, no. 4, pp. 46-51, July/August 2006.

Quinlin, J. R. Induction of Decision Trees, Machine Learning 1:81-106, 1986.

Ross, W.D., *The Right and the Good*, Oxford University Press, Oxford, 1930.

Turing, A.M. Computing machinery and intelligence. Mind, 59, 433-460, 1950.

Waldrop, M. M. A Question of Responsibility. Chap. 11 in Man Made Minds: The Promise of Artificial Intelligence. NY: Walker and Company, 1987. (Reprinted in R. Dejoie et al., eds. Ethical Issues in Information Systems. Boston, MA: Boyd and Fraser, 1991, pp. 260-277.).