2017 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS2017)
5-7 Oct. 2017, Ottawa, Canada

A Decision Making Model for Ethical (Ro)bots

Fahad Alaieri
Management Information Systems / School of Information Studies
Qassim University / University of Ottawa
Quasim 14452, Saudi Arabia / Ottawa, Canada
fahad@uottawa.ca

Andre Vellino
School of Information Studies
University of Ottawa
Ottawa, Canada
avellino@uottawa.ca

***Abstract*— Autonomous bots and robots (we label "(ro)bots"), ranging from shopping assistant chatbots to self-driving cars are already able to make decisions that have ethical consequences. As more such machines make increasingly complex and significant decisions, we need to know that their decisions are trustworthy and ethically justified so that users, manufacturers and law-makers can understand how these decisions are made and which ethical principles were brought to bear in making them. Understanding how such decisions are made is particularly important in the case where a (ro)bot is a self-improving, self-learning type of machine whose choices and decisions are based on past experience, given that they may not be entirely predictable ahead of time or explainable after the fact. This paper presents a model that decomposes the stages of ethical decision making into their elementary components with a view to enabling stakeholders to allocate the responsibility for such choices.**

Keywords*— *Decision making; machine ethics; autonomy; trust, responsibility

I. Introduction (*Heading 1*)

Behaviours exhibited by machines – whether they are deterministic, semi-autonomous or fully autonomous – often have consequences for the well-being of people and therefore have an ethical dimension. Consider, for example, a fully deterministic software system for classification like a spam-filter that incorrectly tags an email. Depending on the tag ("top secret" vs. "spam") and on the sender or recipient, such a software system could have dramatic consequences [1]. This is all the more true in the case of software systems that control the behaviour of mechanical devices such as those actions performed by an automotive collision avoidance system. Even though these deterministic systems make no decisions or choices, they nevertheless can have a significant impact on people's lives and their behaviour has ethical consequences.

The ethical implications of machine behaviour are greater still for machines that make decisions. The making of a decision – a process we describe in greater detail in section 3 – involves a choice from among alternative courses of action. Thus, the ethical relevance of a machine's action that arises from having made a decision – regardless of the actual decision-making process – is all the greater because the machine selected one action from among alternative courses of action. If a machine takes a course of action that has harmful consequences and it was possible for that machine to have chosen a different course of action – especially one that might not have had these harmful consequences – then the act of having made that choice now matters ethically and may even matter legally. For instance, the question of who (or what) is responsible for an action and its consequences is more difficult to determine when alternative courses of action are possible and depends greatly on how the decision was made, which algorithms were used to make it, and whether the choice of action was performed autonomously by the machine.

In this paper we propose a model of machine decision-making that provides a way of better understanding how robots and software bots make decisions in general and ethical decisions in particular. This model lays out the different stages or steps that need to be followed in decision-making and how those decisions, if they are themselves ethical decisions, are governed by ethical principles and other and codes of behaviour such as social values, culture and regulations.

We begin by distinguishing between robots and bots and then explore the characteristics of autonomous ethical (ro)bots and the differences between them and ordinary (ro)bots. We describe ethical principles that ethical (ro)bots follow to make any decisions and discuss the kinds of decision-making processes they execute in section II. The decision-making model that aims to capture the stages of this process is described in section III and we outline its implications in the conclusion.

II. (Ro)bots

It is important to distinguish between "robots" and "bots", even though our decision-making model applies equally to both. A typical definition of a robot is that it is a machine that performs physical actions such as moving and carrying objects and also has some degree of agency and autonomy. On the other hand, a bot, such as a chat-bot or a recommender system does not need to be embodied to make decisions. Kate Darling (MIT Media Lab) says that robots must be viewed as embodied algorithms but that algorithms alone are bots not robots [2]. Another definition provided by Allison Okamura (Stanford CHARM Lab) says that computers (bots) perform information tasks whereas robots perform physical tasks. However, whether or not a machine is embodied or in direct control of physical actions the processes involved in its decision making and their ethical characteristics apply equally to robots and to software agents (bots). We therefore follow Wallach [3] in using the term "(ro)bots" to refer to physical robots and software agents.

978-1-5386-1342-9/17/$31.00 ©2017 IEEE

A. Autonomous (Ro)bots

A key feature in a (ro)bot is whether its ability to make decisions is autonomous, i.e. whether it has the ability to operate without external intervention. From the point of view of its ethical decisions, autonomy is important because it is a necessary condition for ethical agency. While some argue that an autonomous robot cannot be considered truly autonomous unless it makes *all* its decisions without any human intervention, we prefer to say that such robots are not only autonomous but also *independent*. One essential characteristic of an autonomous robot is whether it is able to respond appropriately to a wide variety of situations.

For the purpose of this paper, we distinguish three principal levels of autonomy that a decision-making machine can have: low, partial and full. A (ro)bot with low autonomy relies entirely on human input. Humans are responsible for controlling all the processing stages described in our model below and in such cases it is clear where the responsibility for the action lies. Thus, a machine that requires no external input to make a decision but is only ever able to make one decision could not be said to have a meaningful degree of autonomy. A partially autonomous (ro)bot relies only in part on human input and has the ability to perform most processing stages by itself but may need user input to proceed from one stage to the next, such as a human's acceptance of a choice from among alternative decisions. For example, a collaborative robot like Baxter, which is used to repeatedly perform only very specific tasks, exhibits some degree of autonomy but does not have the ability to make complex decisions that depend on highly variable environmental conditions and is unable to handle unpredictable situations. There again, in those instances where the ethically critical choice-making step is human-controlled, there is no question about where the allocation of responsibility lies. Finally, we say that a (ro)bot is fully autonomous if it relies only on itself, is able to get and analyze information, generate alternatives, evaluate them and commit to a course of action without human intervention. While these three main levels of autonomy could themselves be subdivided into finer categories, the principal distinctions are sufficient for our purpose and the situation of interest in this paper is where a robot is fully autonomous and is required to make decisions.

B. Ordinary (Ro)bots

Not all (ro)bots, even autonomous ones, need to make *ethical* decisions because in many situations there is either no need for ethical principles to guide them or they do not have the capability to make a decision (i.e. chose from among alternatives). For example, low autonomy robots that work on assembly lines and are preprogrammed to install car parts don't have to make ethical decisions even though their behaviour may have harmful consequences: they merely perform routine and repetitive tasks that have predictable inputs, processes and outputs. Assembly robots work in a structured environments and they perform actions automatically but they do not have ability to act autonomously. Automatic (ro)bots perform the same actions, deterministically and predictably given the same initial conditions: they cannot make decisions because they are programmed without any decision making capabilities.

However, in other situations, (ro)bots may have to make ethical decisions because their actions could have consequences that affect humans and it may be possible for them to compute these consequences. For example, autonomous vehicles in the near future will have ability to autonomously transport people from one place to another and, in performing that task, they may need to solve ethical dilemmas. For instance, such a vehicle may have to chose between violating a rule of the road, such as crossing red-light to free the way for an ambulance – and obeying the law but blocking the road for the ambulance vehicle.

C. Autonomous Ethical (Ro)bots

There are three main approaches to governing the ethical behaviour of an autonomous (ro)bot: (i) they can have embedded an explicit set of immutable laws (deontological rules) that control their actions; (ii) they can compute the consequences of their actions and evaluate these consequences according to utilitarian principles (iii) they can be designed with a learning system with the capacity to acquire ethical behaviour from experience (a system of virtue ethics). A combination of each of these approaches is also possible [3].

James Gips [4], for example, introduced a list of requirements that an ethical robot should have: 1) a method to know its environment, 2) a system to calculate actions, 3) a method to predict the results, and 4) a way to calculate the consequences. More recently, Tzafestas [5] maintains that an ethical robot 1) should have an ability to describe every situation in the world, 2) should predict alternative actions, 3) should predict the consequences of the current situation if an action is taken, and 4) have the ability to calculate the goodness of a situation. These requirements are important in autonomous ethical (ro)bots because they are performing actions based on the environment that surrounds them. They need to have the abilities to reaching goals in chosen ways that are suitable for certain situations.

To enable an autonomous (ro)bot to make ethical decisions from experience would require a machine learning method. There are three types of machine learning: supervised learning, unsupervised learning and reinforcement learning [6]. In supervised learning the (ro)bot supervisor teaches the machine to correlate inputs and outputs by using labelled examples so that it may perform these labellings with new inputs and outputs. In unsupervised learning the (ro)bot obtains only the input and, without a supervisor, learns what input patterns are associated with given outputs. In unsupervised learning the (ro)bot obtains the input only and, without a supervisor, learns what input patterns are associated to given outputs. The last type is reinforcement learning in which the (ro)bot learns that a certain goal in a dynamic environment can be achieved by performing a sequence of actions.

IBM Watson, for example, used several of these machine learning techniques to play Jeopardy [7]. It was trained with many questions and known answers (supervised learning) and also used reinforcement learning to learn game strategy. Autonomous (ro)bots are also good candidates to be trained by reinforcement learning applications because they operate in dynamic environments which require a chain of actions that must performed in sequence in order to reach a goal.

D. Ethical Principles

Whether (ro)bots are programmed with explicit deontological ethical rules, given the ability to calculate the ethical consequences of a possible action or trained from experience to develop a disposition to act with virtue, it is critical that their decision-making processes be computationally transparent. If it is possible to determine whether an action is consistent or inconsistent with some generally accepted deontological principles such as a principle of non-harming or a principle of honesty [8], then there is a clear and transparent method of identifying the ethical component of a decision. Similarly a consequentialist approach that calculates the ethical value of a possible action as a function of its consequences can also provide the basis of an explanation for which ethical considerations were involved in a (ro)bot's decision making. However, a virtue-ethics approach, i.e. a system that behaves according to implicit values developed from experience [3] may be more difficult to identify transparently. More complex still is the situation where multiple, hybrid methods are implemented and interact [9].

Whichever one of these ethical frameworks is used to govern an autonomous robot's behaviour, a necessary element for transparency – and hence responsibility and explanations for behaviour – is an understanding of how the machine's decisions are made.

III. Decision Making Model

Following the Sense-Think-Act robotics paradigm in [10, 11, 12] we distinguish ten stages in the decision making processing of a (ro)botic device and identify four of those stages as critical to performing autonomous ethical decisions.

The ten stages in this conceptual model are themselves divided into to two main clusters: the Planning Analysis cluster and the Ethical Guidance cluster (see Figure 1). The main function of planning analysis is to determine whether an action-choice requires an individual, situated ethical analysis or whether it can short-cut to a decision. If it does require ethical analysis, then the four stages in the ethical guidance cluster come into play.

Each stage is best illustrated using example (ro)bots. Consider a delivery robot such as "6D57" by StarShip technologies [13] that delivers packages from point A to point B. Point A could be a pizza store and point B a customer's address. The delivery task is composed of a sequence of subtasks which must be performed in sequence to reach the overall goal – getting the package, navigating on sidewalks or streets to the destination, notifying the customer, delivering the package, and returning home. Another example we use below is that of a chess-playing program that selects a winning move from among alternatives.

Note that, in situations where there are no specificaly *ethical* decisions that a (ro)bot needs to make, it may nevertheless need to chose from among alternative courses of action. Such non-ethical decision making obeys a similar process to the ethical guidance process one described below except that the "objective function" – the quality that the decision aims a maximizing – is different.

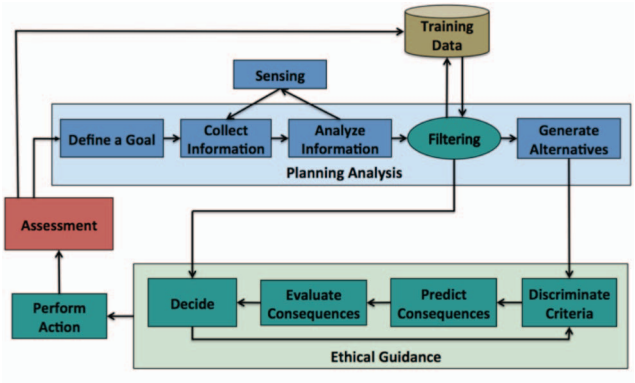

Fig. 1. Ethical Decision Making Model

A. Planning Analysis

1. Defining a Goal.

In our example the main goal of the delivery robot is deliver the pizza to the customer and return home. In this kind of robot the goal is built-in: delivery robots have no other function. A multi-functional robot could have programmable goals and a fully autonomous robot could potentially define its own goals.

2. Collecting Information.

Robotic systems typically acquire information from sensors, networks, telemetry and from humans. The information they need is obtained from sound, light, physical contact and other input methods. Our delivery robot would obtain the customer's address from a data source, determine its own location from a GPS, use its sensors to recognize obstacles and identify paths in the environment and a map to navigate.

3. Information Analysis.

Robotic systems then analyze this input information to create real-world representations of the objects they need to reason about, synthesize and integrate this data to make predictions and execute planning tasks. These two stages - collecting information and analyzing them - are means of "sensing" the environment and may need to be repeated several times to collected the wanted data. In our delivery robot this may mean the identification of obstacles which leads to the dynamic reconfiguration of the trajectory to its destination.

4. Filtering.

Once the information is collected and analyzed the (ro)bot system will then "filter" it. This filtering process occurs in machines that have an experiential learning component, even a rudimentary one. For example, if information collection and analysis has been done once before and a decision has been made on that basis there may be no need to re-compute the decision-making process when similar or identical conditions arise again– it may be simpler to remember the prior decision to retrieve this decision and simply perform the appropriate action. For example, the delivery robot could recognize the entered address if it is for returning customer and navigate the prior path to used by the the robot the last time it delivered a package there.

5. Alternatives Generation.

If the filtering stage indicates that it is necessary to go through the process of making a decision, the (ro)bot then needs to produce a list of alternatives actions from which to select, i.e. to make a decision. The generation of feasible board positions in computer game play is a typical example of this stage. For example, when a computer plays chess it generates dozens alternatives of moves and considers millions of alternative positions that may follow from them using a process similar to the one outlined below for making ethical decisions. In the case of the delivery robot the "alternatives generation" stage would occur if there were alternative routes for reaching its destination; it may, for example, need to select which roads to cross or pick one from among alternative parking spots or decide at which time to notify the customer.

B. Ethical Guidance

The ethical guidance stages govern (ro)botic systems' ethical decision making. These stages embody ethical (deontological and / or utilitarian), principles, including the relevant laws, regulations, cultural values and codes of conduct. Each stage is as follows.

6. Criteria Discrimination.

Given a set of alternative actions from which to choose, the (ro)botic system now should evaluate the generated alternatives according to deontological rules. Some alternatives may need to be eliminated either because they violate ethical codes of conduct or because they are illegal. For instance, in the case of our delivery robot, the need to obey the speed limits (both the upper limits and the lower limits) may preclude certain routes. Similarly parking spots reserved for pregnant women will have to be excluded from possible alternatives.

7. Consequences Prediction.

If the action to be selected needs to be evaluated according to consequentialist principles (stage 8), the implications of each alternative will need to be computed. In the case of the chess-playing (ro)bot, this stage amounts to exploring the search space of possible subsequent board positions that may follow from each alternative. In the case of the delivery robot the choice of route or parking spot may have an effect on the time it takes to achieve its goal, which in turn may have some financial implications for the company or the customer.

8. Consequences Evaluation.

The consequences of each alternative action needs to be evaluated. In the case of a game playing (ro)bot, the consequences are evaluated according to some function that involves the likelihood that this action leads to winning. If there the selection of an action has an ethical dimension, then the evaluation function would have to compute the utility of each alternative according to utilitarian principles. In the simple case of the delivery robot it could be that each rout could have a different likelihood that harm would caused (to itself or to humans) as a result of traffic conditions, the location of schools and perhaps even, prior accident events in the past.

9. Make Decision.

After evaluating the consequences of each alternative action, the (ro)bot will select the act that is optimal with respect to respect to the goal or objective of the overall task and optimal with respect to the ethical. In the case of the delivery robot, this step could take place after an evaluation (in step 8) of the relative ethical risks (possible harm) and benefits (e.g. a timely delivery – and hence a reputational and financial benefit to the company).

10. Performing Actions.

This stage is the one that leads to the commitment to the action. The robot may produce outputs, display information on a monitor, activate an actuator, or communicate with another device or person.

Assessment

The assessment elements govern the learning processes for the training of a (ro)botic systems' ethical decision making. While we cannot cover this stage in detail in this paper, it is a critical one for autonomous, self-learning kinds of (ro)bots. Their learning processes may include an evaluation of the actual actions' consequences compared to the predicted consequences and their input into a data store for training purposes. Such self-learning mechanisms have the virtue of (potentially) improving the quality of the machines' decision making processes and increase the "individuality" of the machines' behaviour: an otherwise identical machine might not behave in the same way under the same circumstances.

This assessment and feedback phase poses an especially important problem from the point of view of allocating responsibilities for action-choices in self-learning machines. An ethically mistaken choice that such an "experienced" machine might make could be the result of prior experiences and any explanation for its behaviour would need to take into account how these experiences influenced the machine's choice.

IV. Conclusions

For ethics-enabled (ro)bots to be trustworthy, their actions must be explainable which in turn requires that the process by which the decisions led to them be transparent. Actions that are the result of a decision making process in which consequences are evaluated using a consequentialist framework or which are assessed by deontological principles, are in particular need of transparency. If ethical decision making processes are opaque, i.e. performed by a "black box" whose inner workings are inaccessible or hard to interpret, then end-users will rightly be concerned about who can be held accountable when mistakes are made. This is particularly true of self-learning (ro)bots whose (virtue) ethical behaviour is learned.

One value of this model is that it provides manufacturers of decision making machines a framework with which to decompose the stages in a system of ethical governance of such decisions. It provides them with guidance concerning the processes need to be monitored (to provide transparency, in case of accidents for example); which phases are responsible for consequence-assessments (e.g. for implementing a

consequentialist ethics) and the role played by deontological "governance" principles. If for example, the Generate Alternatives phase fails to consider alternatives that could have avoided an accident then the question of responsibility for the mistake may reside principally in that module rather than in the prediction or evaluation module. Or, a system that is self-learning might have failed to generate relevant alternative courses of action because of a fault in the Assessment phase.

Our model also provides a framework for legislators and even end-users to understand where ethical responsibilities for machine decisions lie and how to identify the ethical principles that govern these decisions. A machine that is principally driven by a utilitarian calculus in evaluating consequences might be failing to consider alternatives (as above) or it may be predicting consequences incorrectly or evaluating them using the wrong objective function. It could be, for instance, that societal values need to be incorporated either in the objective function against wich consequences are evaluated or elsewhere in the Ethical Guidance cluster of stages.

For example, Microsoft's NLP Twitter chatbot Tay, which was designed to learn from feedback from its conversations with users. Yet, in less than 24 hours after its deployment, it learned it learned from users to tweet in ways that were anti-semitic and racist [14]. This learned behaviour was both unethical and unpredictable because what it learned depended on interactions with its environment and no allowances were made in its design for a potentially "hostile" environment such as malicious humans intent on teaching it to verbally misbehave.

Viewed from the point of view of this decision making model this failure could be attributed to the Criteria Descrimination stage in the Ethical Guidance cluster. A deontological principle that censors discriminatory speech would have prevented the creation of training data that led to undesirable outcomes. Of course, whether such a principle is technologically feasible to implement is a different question.

Finally, as the example of Tay also shows, society at large, including users and legislators will want to know what happened if a robot makes mistake and what a self-learning robot is doing to improve its judgments from experience. It could be that until there are reliable methods of producing human-understandable explanations of learned ethical behaviour, machine-learning methods for the self-improvement of autonomous (ro)bots should be curtailed.

REFERENCES

[1] Vellino, A., Alberts, I. (2016). Assisting the appraisal of e-mail records with automatic classification. *Records Management Journal*, *26*(3), 293-313.

[2] Lafrance, A. (2016, March 22). What Is a Robot? Retrieved January 1, 2017, from https://www.theatlantic.com/technology/archive/2016/03/what-is-a-human/473166/

[3] Wallach, W. and Allen, C., 2009. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

[4] Gips, J. (1991). Towards the Ethical Robot. Presented at the *The Second International Workshop on Human and Machine Cognition Android Epistemology*, Florida.

[5] Tzafestas, S.G. (2016): Roboethics. A Navigating Overview, vol. 79. Springer, Heidelberg

[6] Alpaydın, E. (2010). *Introduction to Machine Learning*. (T. Dietterich, C. Bishop, D. Heckerman, M. Jordan, & M. Kearns, Eds.) (2nd ed.). London: The MIT Press.

[7] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., et al. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3), 59–79

[8] Allen, C., Wallach, W., & Smit, I. (2006). Why Machine Ethics? *Intelligent Systems*, IEEE, 21(4), 12–17. http://doi.org/10.1109/MIS.2006.83

[9] Allen, C., Smit, I., & Wallach, W. (2005). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, 7(3), 149–155. http://doi.org/10.1007/s10676-006-0004-4

[10] Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction. *Journal of Human-Robot Interaction*, 3(2), 74–27. http://doi.org/10.5898/JHRI.3.2.Beer

[11] Vanderelst, D., & Winfield, A. (2017). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*. May 2017 https://doi.org/10.1016/j.cogsys.2017.04.002

[12] Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, *30*(3), 286-297.

[13] Lonsdorf, K. (2017, March 23). Hungry? Call Your Neighborhood Delivery Robot. Retrieved July 3, 2017, from http://www.npr.org/sections/alltechconsidered/2017/03/23/520848983/hungry-call-your-neighborhood-delivery-robot

[14] Mathur, V., Stavrakas, Y., & Singh, S. (2016, December). Intelligence analysis of Tay Twitter bot. *2nd International Conference on Contemporary Computing and Informatics (IC3I),* 2016 (pp. 231-236). IEEE.