The Ethical Characteristics of Autonomous Robots

Fahad Alaieri
falai055@uottawa.ca

André Vellino
avellino@uottawa.ca

Introduction

Autonomous robots such as self-driving cars are already able to make decisions that have ethical consequences. As machines develop a greater capacity to perform a greater number of tasks autonomously, people will increasingly need to trust them to make reliable decisions related to their safety, health, and even their lives. Consequently, autonomous social robots of the future will need to be able to make trustworthy ethical computations regarding the consequences of their actions despite the greater cost and complexity of engineering.

Levels of Autonomy

Low Autonomy

The robot relies entirely on human input. The human is responsible for giving information to the robot, or for analyzing, generating alternatives, deciding, or acting.

Partial autonomy

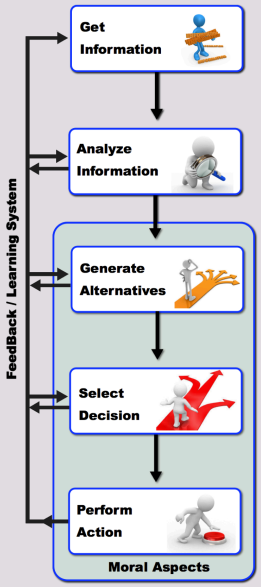
The robot relies only in part on human input and has the ability to perform most processing stage by itself but may need user input to proceed from one stage to the next, such as a human's acceptance of a choice from among alternative decsions.

Full autonomy

The robot relies only on itself, has a high level of intelligence and is able to get and analyze information, generate alternatives and evaluate them and commit to a course of action without human intervention.

Processing Stages

Following [Beer, Jenay, Arthur D. Fisk, and Wendy A. Rogers.] and [R. Parasuraman, T. B. Sheridan, and C. D. Wickens] we distinguish 5 principle stages in the information processing of a robotic device.



Getting Information
Robotic systems may acquire information from sensors, networks, telemetry and from humans. Information may be obtained from sound, light, physical contact and other methods of input.

Information Analysis
Robotic system analyzes this information by categorizing, summarizing, integrating, and predicting.

Alternatives Generation
After analyzing the information the robot should compute a list of alternatives decisions that could be made. Each decision generated at this stage may also contain the computations of the consequences of alternative decisions.

Decision Selection
Evaluation of generated alternatives and their consequences with respect to the goal or objective of the task. After evaluating each alternative, the robot will make a decision to act.

Performing Action
Commitment to the action. The robot may produce outputs, display information on a monitor, activate an actuator, or communicate with another device or person.

Figure 1: Five principle stages in the decision process for a robotic device

Examples

Pepper Humanoid Robot
- Partial autonomy for getting information
- Full autonomy for other stages

Baxter Collaborative Robot
- Low autonomy for getting and analyzing information
- Partial autonomy for generating alternatives and decision selection
- Full autonomy for performing actions

Google Self-Driving Car
- Full autonomy for all processing stages

Relay Hospitality Robot
- Partial autonomy for getting information
- Full autonomy for other stages

Coal Mine Rescue Robot
- Full autonomy for getting information
- Low autonomy for analyzing information, generating alternatives and decision selection
- Partial autonomy for performing action

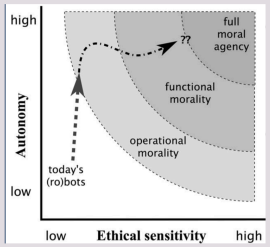
Ethical Characteristics

Moral agency requires the ability to autonomously make the decision to act (see "Alternatives Generation", "Decision Selection" and "Performing Action" robot processing stages) and to make judgements about whether such decisions are either consistent with ethical rules of behaviour or maximize the beneficial outcome of their consequences.

There are two types of Artificial Moral Agents (AMAs):

Type I - is guided by specific ethical rules that cannot be changed and which the AMA must follow. These AMAs exhibit *operational* morality.

Type II - has the ability to modify existing rules and create new rules of behaviour based on what it learns from its surroundings.



- Operational morality: a robot's actions are pre-programmed, deterministic and predictable.

- Functional morality: a robot's actions not entirely predictable because the robot has the ability to act without human control and its choices are determined by many input variables.

- Full morality: a robot's actions can be performed fully autonomously and its decision-making processes are not necessarily known (self-learning autonomous robot).

Figure 2. Two Dimensions of AMA Development (Wallach & Allen, 2009)

For people to trust autonomous robots that make decisions with ethical consequences they need to be programmed to obey ethical principles whose application is predictable and traceable (Type I). However, Type II AMAs that have ability to modify the method by which they generate alternatives and calculate the consequences of their possible future actions may not be entirely predictable. If a robot can self-modify its decision procedures, its behaviour may become non-deterministic and unpredictable, and how it came to make a choice may be complex, and hard to explain.

It may also not be possible to for a fully autonomous robot, even one that is not self-learning, to both apply predictable rule-based ethical principles that people can trust and make optimal decisions based on the calculation of their consequences.

Selected References

- R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation.," IEEE Transactions on Systems, Man and Cybernetics, Part A Systems and Humans, vol. 30, no. 3, pp. 286–297, May 2000.
- B. T. Clough, "Metrics, Schmetrics! How The Heck Do You Determine A UAV's Autonomy Anyway?," presented at the Proceedings of the Performance Metrics for Intelligent Systems Workshop PerMIS -, Gaithersburg, MD, 2002.
- M. Nagenborg, "Artificial moral agents: an intercultural perspective ," International Review of Information Ethics, vol. 7, pp. 1–6, 2007.
- W. Wallach and C. Allen, Moral Machines Teaching Robots Right from Wrong Oxford University Press, 2009.
- J. M. Beer, A. D. Fisk, and W. A. Rogers, "Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction," Journal of Human-Robot Interaction, vol. 3, no. 2, pp. 74–27, Jun. 2014.