Abstract

Una delle principali sfide tecnologiche e scientifiche nello sviluppo di macchine e robot autonomi consiste nel garantire il loro comportamento etico e sicuro nei confronti degli esseri umani. Nel caso di macchine autonome, infatti, l'operatore umano non è presente, quindi la complessità del rischio deve essere interamente gestita dall'intelligenza artificiale delle macchine e dai loro sistemi di decisione, che devono essere concepiti e progettati in modo da garantire un comportamento sicuro ed etico. In questo lavoro verrà proposto un possibile approccio per lo sviluppo di sistemi decisionali per macchine autonome, basato sulla definizione di criteri e principi etici generali. Questi principi riguardano la necessità di evitare o ridurre al minimo il verificarsi di danni agli esseri umani, durante l'esecuzione del compito per cui la macchina è stata progettata.

In questo ambito si possono introdurre quattro problemi fondamentali:

1. Primo Problema: identificare leggi o principi etici applicabili alle macchine

2. Secondo Problema: incorporare l'etica nella macchina

3. Terzo Problema: definire il grado di interazione uomo-macchina

4. Quarto Problema: prevenire l'utilizzo scorretto della macchina.

L'attività di ricerca condotta nell'ambito di questo dottorato si è concentrata principalmente sul Primo e Secondo Problema, con specifico riferimento agli aspetti di sicurezza. Per quanto riguarda il Primo Problema, lo scopo principale di questo lavoro è garantire che una macchina autonoma agisca in modo sicuro, ovvero:

• Nessun danno venga arrecato agli esseri umani circostanti (principio etico di non-maleficenza)

• Nel caso in cui un essere umano si avvicini a una potenziale fonte di danno, la macchina deve agire in modo tale da ridurre al minimo tali danni con la migliore azione possibile e disponibile (principio etico di non-inazione)

e, quando possibile e non in conflitto con i principi di cui sopra:

• La macchina deve agire in modo da preservare la propria integrità (autoconservazione).

Per quanto riguarda il Secondo Problema, una versione semplificata dei principi etici sopra riportati è stata utilizzata per costruire un modello matematico di un sistema di decisione sicuro basato sulla teoria dei giochi. Limitandosi alla sicurezza e non estendendo all'intera sfera generale dell'etica, è possibile adottare dei criteri ben definiti nell'assicurare che il comportamento della macchina non arrechi alcun danno agli esseri umani, ad esempio:

• Verificando sempre che la macchina mantenga una distanza di sicurezza adeguata a una certa velocità di funzionamento

• Assicurandosi sempre che, entro un certo intervallo di tempo, la macchina possa rilevare la distanza tra un essere umano e la posizione di un potenziale danno.

One of the major technological and scientific challenges in developing autonomous machines and robots is to ensure their ethical and safe behavior towards human beings. When dealing with autonomous machines the human operator is not present, so that the overall risk complexity has to be addressed to machine artificial intelligence and decision-making systems, which must be conceived and designed in order to ensure a safe and ethical behaviour. In this work a possible approach for the development of decision-making systems for autonomous machines will be proposed, based on the definition of general ethical criteria and principles. These principles concern the need to avoid or minimize the occurrence of harm for human beings, during the execution of the task the machine has been designed for.

Within this scope, four fundamental problems can be introduced:

1. First Problem: Machine Ethics Principles or Laws Identification
2. Second Problem: Incorporating Ethics in the Machine
3. Third Problem: Human-Machine Interaction Degree Definition
4. Fourth Problem: Machine Misdirection Avoidance.

This Ph.D. research activity has been mainly focused on First and Second Problems, with specific reference to safety aspects. Regarding First Problem, main scope of this work is on ensuring that an autonomous machine will act in a safe way, that is:

• No harm is issued for surrounding human beings (non maleficence ethical principle)

• In case a human being approaching a potential source of harm, the machine must act in such a way to minimize such harm with the best possible and available action (non-inaction ethical principle)

and, when possible and not conflicting with above principles:

• The machine must act in such a way to preserve its own integrity (self-preservation).

Concerning Second Problem, the simplified version of some ethical principles reported above has been used to build a mathematical model of a safe decision system based on a game theoretical approach. When dealing just with safety and not with general ethics, it is possible to adopt some well-defined criteria in ensuring the machine behaviour is not issuing any harms towards human beings, such as:

• Always ensure the machine is keeping a proper safety distance at a certain operating velocity

• Always ensure that, within a certain range, the machine can detect the distance between a human being and the location of a potential harm.