Novas fronteiras da ética na Informática: Inteligência Artificial e Agentes Autónomos

João Campinhos e Pedro Durães

Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa,
Quinta da Torre, 2829 – 516 Caparica, Portugal
j.campinhos@campus.fct.unl.pt,
p.duraes@campus.fct.unl.pt
http://http://www.fct.unl.pt

Resumo Este documento visa analisar o tema da inteligência artificial e agentes autónomos no âmbito dos aspectos sócio-profissionais da informática. Iremos abordar não só a forma como a inteligência artificial pode ajudar a humanidade, mas também os problemas éticos e riscos que levanta. Com o desenvolvimento de agentes autónomos como nos carros sem condutor, deparamo-nos com inúmeros riscos que precisam de ser ponderados, para que o desenvolvimento da inteligência artificial não seja prejudicial.

Keywords: As palavras chave são aqui!

1 Introdução

A inteligência artificial está muito associada a robôs pois este é um tema bastante abordado no cinema. Geralmente nesses filmes, os engenheiros que programam estes agentes autónomos criam algo maior que eles próprios e que deixam de poder controlar. Mas isto é ficção, e o nosso objectivo com este documento é tentar aproximar este tema da realidade, pois apesar de ainda estarmos numa fase bastante embrionária no desenvolvimento de agentes autónomos, é uma área potencialmente perigosa, e cabe-nos a nós, engenheiros e futuros engenheiros informáticos, com um desenvolvimento responsável fazer com que o mau da inteligência artificial nunca venha ao de cima.

Existe ainda uma certa relutância e negação quando é abordado este tema, mas na verdade cada vez mais pessoas importantes na área estão a levá-lo a sério, como é o caso do Elon Musk, cofundador do Paypal e Tesla Motors e fundador da empresa de exploração espacial SpaceX, que doou 10 milhões de dólares para o instituto Future of Life Institute, numa tentativa de promover um desenvolvimento da inteligência artificial e agentes autónomos estritamente beneficiais para a humanidade.

2 Contexto

TODO: FALAR SOBRE CONSCIÊNCIA E ROBÔS E ASSIM

Em 1942, foram introduzidas numa obra de ficção de Isaac Asimov as famosas três leis da robótica:

1. Um robô não pode ferir um ser humano ou, por inacção, permitir que um ser humano sofra algum mal.
2. Um robô deve obedecer as ordens que lhe sejam dadas por seres humanos excepto nos casos em que tais ordens entrem em conflito com a Primeira Lei.
3. Um robô deve proteger sua própria existência desde que tal protecção não entre em conflito com a Primeira ou Segunda Leis.

Estas leis começam por levantar um problema importantíssimo na inteligência artificial: o mal que um robô pode fazer ao ser humano. É claro que, em 1942 esse não era um problema, mas na actualidade, e com o desenvolvimento da inteligência artificial, começamos a ter de pensar nestes casos, e analisar o problema de forma objectiva.

À primeira vista, as três leis de Isaac Asimov parecem ser bastante estritas e claras, e ter um robô a respeitá-las parece ser bastante seguro. O que é certo é uma analise mais cuidada levanta algumas ambiguidades importantes. A começar pela própria definição de robô e ser humano, que pode ser mal interpretada pelo robô fazendo com que não respeite as três leis. Outro problema reside no facto do robô poder quebrar uma lei sem se aperceber. O que acontece nessa situação? E existem mais ambiguidades, mas só por aqui podemos ver a complexidade da inteligência artificial, e o impacto que qualquer interpretação errada pode ter.

Referências