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.

3 Agentes de Super-inteligência Artificial

Falar um bocado disto

É importante fazermos a distinção no que a super-inteligência artificial diz respeito

3.1 Super-inteligência de velocidade

Podemos comparar a super-inteligência de velocidade como um supercomputador comparado com um computador normal. Neste caso, estamos perante um agente que consegue pensar muito mais rápido que um humano, mas que pensa da mesma forma.

3.2 Super-inteligência de qualidade

Aqui é onde um agente SIA brilha em relação aos humanos. Não só vai pensar muito mais rápido que um humano como também vai pensar de forma muito mais avançada, impossível aos humanos de compreender. A diferença nas super-inteligências é fácil de ver se imaginarmos este cenário: Um chimpanzé a pensar a uma velocidade superior à dos humanos não é suficiente para que seja superior. A evolução fez com que os humanos desenvolvessem capacidades cognitivas que

os chimpanzés simplesmente não conseguem percebem, não importa a velocidade do seu cérebro.

O mesmo se passa quando relacionamos um agente SIA com o ser humano, mas a um nível bastante superior, pois esperamos que um SIA possa evoluir muito mais depressa que um humano, pelo que a diferença entre um SIA e um humano possa ser cinco vezes ou mais superior que a diferença entre um humano e uma galinha.

E com isto é bastante importante ressalvar que é impossível sabermos as consequências que um agente SIA irá ter nas nossas vidas e na humanidade. Estamos perante uma bomba relógio que não compreendemos e que só podemos tentar fazer para que expluda da forma mais inócua possível. Segundo Nick Bostrom, filósofo da universidade de Oxford, podemos dividir as consequências de um agente SIA em duas partes distintas. Ou a extinção da humanidade, ou a sua imortalidade.

Até agora, 99,9% das espécies que viveram na terra extinguiram-se. É natural que os humanos sigam o mesmo caminho. A não ser que isso não aconteça. Bostrom acredita que ainda não houve nada no planeta terra que fosse inteligente o suficiente para atingir a imortalidade, mas que tal não é impossível com um agente SIA. O impacto será tão elevado que ou caímos para um lado do espectro ou para outro.

3.3 Controlo do poder

Como podemos constatar, quando nos referimos a um agente com super-inteligência estamos a falar de algo com um poder inimaginável. O que acontece se for um grupo com más intenções o primeiro a desenvolver um agente SIA?

Na verdade, a comunidade científica não está muito preocupada com este cenário, pois na verdade o problema não está nas intenções dos criadores do agente SIA mas sim no facto desse grupo ter apressado o seu desenvolvimento sem utilizar uma abordagem responsável, levando à perda do controlo sobre o agente SIA. Existe sempre o problema de um agente SIA ser criado por um grupo com fins maliciosos, mas qualquer grupo vai ter o mesmo problema em controlar o agente super-inteligente, seja esse grupo bom ou mau.

FALAR DOS BIONICOS

3.4 Destruição mundial e consciência

A ficção habituou-nos a este tipo de cenário: Um agente super-inteligente malévolo decide destruir a humanidade. Este cenário retratado no cinema não faz muito sentido na realidade. O bem e o mal são conceitos humanos, e pensar que algo não humano pode perceber estes conceitos é chamado de Antropomorfismo. Como já vimos anteriormente, um SIA destrutivo é possível, mas não por ser malicioso.

Isto levanta outro grande tópico relacionado com a inteligência artificial: a consciência. Será que vamos chegar a ter uma inteligência artificial que consiga rir e sentir as mesmas emoções que nós, ou simplesmente simular essas emoções

e essa consciência? Esta questão tem sido explorada dando origem a argumentos como o The Chinese Room **FOOTNOTE!**

Se efectivamente chegarmos a ter humanos artificiais como Kurzweil acredita, será que vamos poder desliga-los como faríamos com um computador ou será visto como um assassinato? Esta é mais uma questão relacionada com a ética quem precisa de ser resolvida.

O problema reside não no facto de se poder criar uma inteligência artificial cujo objectivo é acabar com a humanidade, mas sim no criar uma inteligência artificial criada para cumprir determinada tarefa e que, para a cumprir, acabe com a humanidade ou destrua o mundo como nós os conhecemos. Parece pouco provável mas não é tão diferente do homem matar um animal para comer. O objectivo não é matar o animal mas sim consumir os nutrientes necessários para sobreviver. O animal morrer é efeito colateral do nosso objectivo.

TURRY PARA BAIXO

Referências