11.6 DIFICULDADES DA PROGRAMAÇÃO DA ÉTICA EM ROBÔS

Muita coisa tem sido dita sobre os futuros desenvolvimentos em robótica. Uma possibilidade é que os robôs terão uma dimensão ética. Conforme veremos, isso não envolve necessariamente livre arbítrio ou consciência nos robôs.

Uma descrição interessante de como a ética pode ser programada em robôs foi apresentada em um artigo intitulado "Towards Machine Ethics" (Rumo à Ética na Máquina), escrito por Michael Anderson, Susan Leigh Anderson e Chris Armen em 2004. Eles escrevem que, "Ao contrário da invasão em computadores, das questões de propriedade de software, das questões de privacidade e de outros tópicos normalmente relacionados à ética na *computação*, a ética na *máquina* trata das consequências do comportamento de máquinas com respeito a usuários humanos e a outras máquinas".[9]

Eles argumentam que, à medida que se atribui maior responsabilidade às máquinas, é apropriado que se cobre uma medida igual de responsabilização por parte delas. Isso, naturalmente, requer que elas disponham de um processo pelo qual possam tomar decisões éticas e pelas quais possam ser responsabilizadas. Por meio de exemplos de duas teorias éticas, os autores mostram como tal processo pode ser programado em robôs.

O primeiro exemplo é o Utilitarismo dos Atos, uma teoria formulada por Jeremy Bentham no final do século XVIII. Essa teoria afirma que ético é aquilo que promove a felicidade. Bentham considerava a felicidade o excedente de prazer sobre a dor para aqueles afetados por qualquer ação conhecida (veja o Capítulo 3).

Os autores decidiram formular um algoritmo que calcula a ação, dentre todas as alternativas, que produz o maior prazer 'líquido'. Eles escrevem que isso "requer como entrada o número de pessoas afetadas e, para cada pessoa, a intensidade do prazer/desprazer (por exemplo, em uma escala de 2 a –2), a duração do prazer (por exemplo, em dias) e a probabilidade de ocorrência desse prazer/desprazer para cada ação possível. Para cada pessoa, o algoritmo simplesmente calcula o produto da intensidade, da duração e da probabilidade, para obter o prazer líquido para cada pessoa. Em seguida, ele soma o prazer líquido individual para obter o Prazer Líquido Total... Esse cálculo seria realizado para cada ação alternativa. A ação com o mais alto Prazer Líquido Total é a ação correta".[10]

Os autores salientam que o Utilitarismo dos Atos tem sido criticado por violar os direitos de uma pessoa, ao sacrificar uma pessoa pelo bem maior. Assim, eles dão outro exemplo para mostrar como essa deficiência pode ser corrigida utilizando as teorias de W. D. Ross e John Rawls.

A teoria de Ross é baseada em dever, em vez de seguir a orientação de Bentham voltada para as consequências. Ross propõe sete obrigações de evidência aparente (*prima facie*): fidelidade, reparação, gratidão, justiça, beneficência, não maleficência e

aperfeiçoamento pessoal. Infelizmente, Ross não fornece um meio para determinar que obrigação é a mais forte. Nesse ponto, os autores sugerem a abordagem do "equilíbrio reflexivo" de Raws.[11] Ela envolve: a) considerar as atribuições de peso possíveis de todas as obrigações pertinentes; b) testá-las por meio de nossas intuições com respeito a casos particulares; c) revisar as atribuições de peso de modo a refletir nossas intuições; d) testá-las novamente. Em suma, os autores dizem que "em vez de calcular um único valor com base apenas em prazer/desprazer, devemos calcular a soma de até sete valores, dependendo do número de obrigações de Ross pertinentes à ação em particular. O valor para cada uma dessas obrigações poderia ser calculado tal como o Utilitarismo dos Atos Hedonístico, como um produto da Intensidade, Duração e Probabilidade."[12]

O processo descrito parece cansativo se for feito por um ser humano, mas seria bem adequado à computação de máquina. Naturalmente, o algoritmo e a programação teriam a contribuição de programadores humanos e, portanto, a antiga regra de que nenhum programa pode ser melhor que seu programador seria mantida.

Assim, ficamos com a seguinte questão: embora pareça possível programar uma máquina para tomar decisões éticas, tais tomadas de decisão poderiam ser descritas como autônomas?

Trecho extraído do livro:
ETICA NA COMPUTAÇAO: *uma abordagem baseada em casos*
Robert N. Barger, Editora LTC 2011