Da maquinaria da moral à moral da máquina

Luís Moniz Pereira[1]

Desde sempre os seres humanos se aperceberam dos riscos associados tanto ao conhecimento, quanto às tecnologias a ele associadas. Não só na mitologia grega se encontram os sinais e os alertas para esses perigos, mas também nos mitos fundadores das religiões judaico-cristãs. Todavia, tais alertas e receios nunca fizeram tanto sentido quanto o que fazem hoje em dia. Tal resulta da emergência de máquinas capazes de reclamar para si funções cognitivas que até há pouco eram desempenhadas exclusivamente por seres humanos. A revolução cognitiva, propiciada pelo desenvolvimento da IA, além dos problemas técnicos associados ao seu desenho e conceptualização, levanta problemas de ordem social e económica com impacto directo na humanidade em geral, e nos seus grupos constituintes. Daí que a sua abordagem do ponto de vista moral seja urgente e imperiosa. Os problemas éticos a considerar são de duas ordens: Por um lado os que estão associados ao tipo de sociedade que queremos promover através automação, complexificação e ao poder de tratamento de dados hoje em dia disponíveis; por outro, como programar máquinas para tomada de decisão de acordo com princípios morais aceitáveis pelos seres humanos que com elas partilham o conhecimento e a acção.[2]

No período Helenístico (323-31 a.C.), Hero de Alexandria e outros brilhantes engenheiros gregos vinham congeminando uma diversidade de máquinas, movidas quer hidráulica quer pneumaticamente. Os Gregos reconheciam que autómatos e outros artefactos com formas naturais – imaginários ou reais – poderiam ser tanto inofensivos quanto perigosos. Lograriam também ser usados para trabalho, sexo, espectáculo ou religião, ou para infligir dor ou morte. Claramente, a biotecnologia, real e imaginária, fascinava já os Antigos.[3]

Os actuais algoritmos de aprendizagem profunda permitem que os computadores da IA extraiam padrões em dados vastos, que os extrapolem para situações novas e tomem decisões sem qualquer orientação humana. Inevitavelmente, as entidades da IA desenvolverão a capacidade de se interrogar, e responderão às questões que venham a descobrir. Os computadores de hoje já mostraram saber desenvolver o altruísmo, mas também o ludibriar outrem por conta própria. Faz pois todo o sentido a questão: Porquê uma ética para máquinas?

- Porque os agentes computacionais se tornaram mais sofisticados, mais autónomos, actuam em grupo, e formam populações que incluem humanos.
- Estão a ser desenvolvidos numa variedade de domínios, onde questões complexas de responsabilidade exigem maior atenção, nomeadamente em situações de escolha ética.
- Uma vez que a respectiva autonomia está a aumentar, o requisito de que funcionem com responsabilidade, eticamente, e de modo seguro, é uma preocupação crescente.

A deliberação autónoma e criteriosa reclama regras e princípios de natureza moral aplicáveis à relação entre máquinas, à relação entre máquinas e seres humanos, e às consequências e

---

[1] Laboratory for Computer Science and Informatics (NOVA-LINCS), Departamento de Informática, Faculdade de Ciência e Tecnologia, Universidade Nova de Lisboa.

[2] Para o desenvolvimento desta temática veja-se o nosso recente livro *Da Moral da Máquina à Maquinaria da Moral* de Luís Moniz Pereira e António Barata Lopes, Lisboa: NOVA.FCT Editorial, 2020. Também publicado em inglês L. M. Pereira, A. Lopes, ***Machine Ethics: From Machine Morals to the Machinery of Morality***, na série Studies in Applied Philosophy, Epistemology and Rational Ethics (SAPERE, volume 53), Cham: Springer Nature, Switzerland AG, **2020**.

[3] Consultámos *Gods and Robots – Myths, Machines, and Ancient Dreams of Technology* de Adrienne Mayor, Princeton, NJ: Princeton U. Press, 2018, e ainda *A Brief Guide to the Greek Myth*s de Stephen P. Kershaw, London: Constable & Robinson Ltd., 2007.

resultados da entrada destas máquinas no mundo do trabalho e na sociedade em geral. O presente estado de desenvolvimento da IA, tanto na sua capacidade de elucidação dos processos cognitivos emergentes na evolução, quanto na sua aptidão tecnológica para a concepção e produção de programas informáticos e artefactos inteligentes, constitui o maior desafio intelectual do nosso tempo. A complexidade destas questões ilustra-se sinteticamente com um esquema, mais à frente – designado por "O carrossel da maquinaria ética". Em síntese, estamos numa encruzilhada, a da IA, da ética das máquinas, e do seu choque social.

O tópico da moral tem dois grandes domínios. O primeiro designado por "cognitivo", ou seja, em torno da necessidade de se esclarecer como é que pensamos em termos morais. Para ter um comportamento moral é preciso equacionar possibilidades: devo eu comportar-me deste modo ou comportar-me de um outro? É preciso avaliar os vários cenários, as várias hipóteses. É preciso comparar essas hipóteses para ver quais são as mais desejáveis, quais são as respectivas consequências, quais são os seus efeitos laterais. Esta capacidade é essencial para a vivência em sociedade, para o domínio do "colectivo".
Estudei certas capacidades cognitivas para depois ver se eram promotoras de cooperação moral, numa população de seres informáticos, i.e. de programas de computador a conviverem entre si. Considerando que um programa é um conjunto de estratégias definidas por regras; ou seja, numa dada situação, um programa desenvolve uma certa acção ditada pela sua estratégia vigente, e em que os outros programas têm igualmente acções ditadas pelas respectivas regras estratégicas. É como se fossem agentes convivendo em conjunto, cada um com opções de objectivo possivelmente diferentes. Estuda-se o "se", e o "como" essa população vai poder evoluir num bom sentido gregário, e se esse sentido é estável, i.e., se se mantém no tempo[4].

Um instrumento muito importante para esta investigação é a chamada Teoria dos Jogos Evolucionários (*Evolutionary Game Theory* [5]), que consiste em ver como, num jogo com regras bem definidas, uma população evolui por aprendizagem social. A sociedade rege-se por um conjunto de preceitos de funcionamento em grupo, digamos regras de um jogo em que é permitido fazer certas coisas, mas não outras. O jogo indica quais os ganhos ou perdas de cada interveniente em cada lance, consoante o modo como joga. A aprendizagem social consiste em que um jogador passe a imitar a estratégia de um outro cujos resultados indicam ter tido maior sucesso. Definidas certas regras, como é que evolui o jogo social? Aqui poderíamos entrar pelo campo da ideologia, mas não iremos tão longe. Estamos ainda a estudar a viabilidade da moral. Partimos do princípio de que a moral é evolucionária, que se desenvolveu com a nossa espécie. À medida que nos transformámos, por centenas de milhar de anos, fomos aperfeiçoando as regras de convivência e aprimorando as nossas próprias capacidades intelectuais e saber como usar essas regras de convívio. Nem sempre convenientemente, ou seja, as regras sociais deveriam ser de forma a todos beneficiarmos, embora haja sempre a tentação de alguns quererem, iníqua e injustamente, mais que os outros – de usufruírem das vantagens sem pagar para os custos. Trata-se do problema essencial da cooperação: como é que esta se torna possível e, simultaneamente, como se mantém sob controlo os que dela querem abusar. Para a nossa espécie chegar onde chegou até hoje, a própria evolução teve que nos ir selecionando em termos de uma moral de convivência gregariamente proveitosa.

O problema do progresso da cooperação e do surgimento de um comportamento colectivo, que atravessam disciplinas tão diversas como a Economia, a Física, a Biologia, a Psicologia, as Ciências Políticas, as Ciências Cognitivas e a Computação, é ainda um dos maiores desafios interdisciplinares que a ciência enfrenta hoje. Técnicas matemáticas, e de simulação, da Teoria dos Jogos Evolutivos, têm vindo a mostrar-se úteis para estudar tais

[4] Se desejar aprofundar o tema poderá visitar a página do autor em https://userweb.fct.unl.pt//~lmp/publications/Biblio.html
[5] https://en.wikipedia.org/wiki/Evolutionary_game_theory

matérias. A fim de se compreender melhor os mecanismos evolutivos que promovem e mantêm o comportamento cooperativo em variadas sociedades, é importante levar em conta a complexidade intrínseca dos indivíduos participantes, ou seja, os seus intrincados processos cognitivos na tomada de decisão. O resultado de muitas interacções sociais e económicas é definido, não apenas pelas previsões que os indivíduos fazem quanto ao comportamento e às intenções de outros indivíduos, mas ainda pelo mecanismo cognitivo que os outros adoptam para tomar as suas próprias decisões.

A investigação, baseada em modelos matemáticos abstractos para esse efeito, mostrou que a maneira como o processo de decisão é modelado tem uma influência variada no equilíbrio a alcançar nas dinâmicas de colaboração colectiva. São abundantes as provas mostrando que os humanos (e muitas outras espécies) possuem habilidades cognitivas complexas: teoria-da-mente; reconhecimento de intenções; raciocínios hipotético, contrafactual e reactivo; orientação emocional; aprendizagem; preferências; compromissos; e moralidade. Para melhor compreender como todos estes mecanismos possibilitam a cooperação, é preciso que sejam modelados adentro do contexto dos processos evolutivos. Por outras palavras, devemos procurar entender como é que os sistemas cognitivos usados para explicar o comportamento humano – desenvolvidos com sucesso pela Inteligência Artificial e Ciências Cognitivas – lidam com a teoria evolutiva de Darwin, e por essa via se perceba e justifique o seu aparecimento em termos da existência de uma dinâmica de cooperação, ou da ausência dela.

Deve enfatizar-se, no entanto, que estamos perante *Terra Incognita*. Há todo um continente por explorar, cujos contornos apenas vislumbramos. Não sabemos ainda o bastante sobre a nossa própria moral, nem os nossos conhecimentos são suficientemente exactos para a poderem ser programados em máquinas. Na verdade, existem várias teorias éticas, antagonistas entre si, mas que também se complementam. A Filosofia e a Jurisprudência estudam a Ética, que é a problemática de definir um sistema de valores articulado em princípios. Cada ética em particular é o substracto que serve de suporte às normas e legislação que justificam, em cada contexto, as regras específicas que irão ser aplicadas e usadas no terreno essa ética. Como resultado, e dependendo das culturas e das circunstâncias, chega-se às regras morais. Em cada contexto, parte-se de princípios éticos abstractos para se chegar a regras morais concretas. Na prática, uma moralidade, um conjunto de regras morais, resulta de uma combinação histórica, contextual e filosófica de teorias éticas que foram evoluindo no tempo.

O *Carrossel da Maquinaria Ética*, abaixo, resume de certa maneira a complexidade da problemática da maquinaria moral. No carrossel central querem identificar-se aqueles factores que dizem respeito ao que fazer, ou como agir. "O que fazer" está rodeado de outros tantos carrosséis, cada um tendo a ver com o uso ético das máquinas.

Sobre o uso ético, já ouvimos falar de *fake-news* e de algoritmos que influenciam eleições: trata-se de um mau uso das máquinas, que deverá estar sujeito a regras morais. Por exemplo, é uma prática negativa, imoral, um programa fazer-se passar por um ser humano. Há, claro, outros exemplos de usos imorais das máquinas. Dos mais tenebrosos serão os drones com capacidade autónoma para matar indivíduos. É forçoso recordar, portanto, que o desregramento do uso cada vez mais se subordina à capacidade da própria máquina, justamente porque esta tem cada vez maior autonomia, o que, por consequência, amplifica as questões do seu uso moral.

Isso permite-nos pensar que as máquinas deveriam também proteger-nos do seu uso não-ético por parte dos humanos. Suponhamos que alguém manda um programa executar uma instrução para agir causando prejuízo a seres humanos – o próprio programa poderia recusar-se a fazer tal.[6] Será essa a segunda razão pela qual precisamos de introduzir moral nas

[6] Trata-se da 1ª das *Três Leis da Robótica* idealizadas por Isaac Asimov, condensadas na *Lei Zero*: Um robot não pode causar mal à humanidade ou, por omissão, permitir que a humanidade sofra algum mal.

máquinas, para que não cumpram tudo o que forem meramente programadas para fazer. Não queremos que a máquina esteja em situação do simplesmente "fiz isso porque me mandaram". Uma metáfora para a posição dos criminosos de guerra nazis em Nuremberga, dizendo "apenas cumpri ordens, fiz o que me mandaram," como se não tivessem sentido crítico e não pudessem desobedecer a ordens. Coloca-se o desafio de construir máquinas capazes de desobedecer quando justificado.

Outro círculo do carrossel é o dos Valores Humanos. No fundo, pretendemos dar às máquinas os nossos valores, porque elas vão conviver connosco. Estas terão que estar eticamente conciliadas com a população onde se encontrarem.

O carrossel da maquinaria ética

Noutro círculo do carrossel assinala-se a Legislação porque, no fim do processo, tudo terá que ser traduzido em leis, normas e padrões, sobre o que é permitido ou proibido. Tal como os carros têm regulamentos quanto à poluição, também as máquinas terão que obedecer a certos critérios, aprovados por entidade capacitada para o fazer. Muitas vezes se pergunta quem é responsável se um carro sem condutor atropelar um peão quando pudesse não o ter feito? O dono, o fabricante? Mas omite-se falar no Legislador. Contudo, alguém teve que dizer "este carro sem condutor pode circular". Terá que ser um governo a legislar sobre quais os testes que um carro sem condutor deve superar. Caso se verifique que tais testes não foram suficientemente exaustivos, a entidade que autorizou a circulação daqueles veículos será também responsável.

Um outro círculo é o dos Assuntos Técnicos. Tudo envolve sempre a parte de construção efectiva das máquinas, para o que quer que seja. Nem tudo é tecnicamente possível. Por exemplo, ainda não soubemos pensar os termos da prova informática de que uma máquina não vá fazer coisas eticamente incorrectas. Já para não falar nos casos em que um qualquer 'hacker' possa entrar no sistema e obrigar a máquina a executar coisas erradas. Trata-se de um problema de segurança a ser resolvido de modo técnico.

Por fim, e não menos importante, são os Impactes Sociais das máquinas com autonomia. Quando nos referimos a máquinas estamos a falar quer em robôs quer em 'software'. Este último é bem mais perigoso pois espalha-se e reproduz-se facilmente para qualquer lugar do mundo. Quanto ao robô, será muito mais difícil de reproduzir, implica um custo muito maior, e traz as limitações materiais inerentes à posse de um corpo volumétrico. No que toca ao impacte social, é expectável que a breve trecho tenhamos robôs a cozinhar hambúrgueres e a servir-nos à mesa, com as implicações daí resultantes para o mercado de trabalho. Embora para essas tarefas não se exija muita inteligência, os desafios inerentes a uma coordenação fina olho-cérebro-mão não são de negligenciar. Algo que as máquinas ainda não têm tanto quanto os humanos, mas também nessa frente os robôs estão a avançar muito depressa. No que toca ao 'software', a questão é mais preocupante porque, no fundo, os programas estão a alcançar níveis cognitivos que até agora eram nosso monopólio. Daí que as pessoas se sintam muito mais preocupadas. Ou seja, havia até agora coisas que só um ser humano sabia fazer. Porém, pouco a pouco, as máquinas começaram a jogar xadrez, a fazer diagnósticos médicos, etc., e cada vez mais irão desempenhar actividades mentais mais sofisticadas, e irão também, sucessivamente, aprendendo a fazê-lo.

Esta porta que se abre cria, pela primeira vez, uma competição aberta com os humanos. Uma concorrência que poderá fazer com que – dependendo da organização social, da ideologia, e da política – os humanos venham a ser substituídos por máquinas. Porque, para fazer uma mesma coisa, haverá instrumentos mais baratos do que o humano. Então, tornando-se o humano dispensável, os salários irão diminuir e os donos das máquinas irão ficar cada vez mais ricos. O presente hiato de riqueza, que vem aumentando e indica que os ricos estão cada vez mais ricos e os pobres cada vez mais pobres, é uma brecha que a IA está já a cavar, e virá a ampliar ainda mais. A ponto que se exigirá um novo contrato social, sob pena de um cataclismo. A maneira como funcionamos em termos de capital e trabalho, e o modo como as duas coisas se equacionam, terão que ser completamente reformulados. Corre-se o risco de, caso tal não aconteça, que as assimetrias quanto à riqueza façam com que, mais tarde ou mais cedo, ocorra uma grande revolta, insurreição, e desagregação social. Tal acontecerá quando o sistema de castas induzido pelos avanços da IA gerar a sua própria implosão.

Actualmente, já temos robôs em hospitais, temos drones que voam por si, temos lanchas aquáticas autónomas, temos carros sem condutor, e temos até jogos morais interactivos, susceptíveis de ensinar moral. Num jogo que desenvolvi um robô vai salvar uma princesa, para o que combina várias aproximações éticas[7]. Inclusive, tal exemplifica-nos que, na prática, a moral não é uma só, mas antes uma combinação múltipla. Nós próprios não seguimos exclusivamente a moral do cavaleiro andante, ou a moral utilitária, ou a moral kantiana, ou a moral de Gandhi. A nossa ética é uma mistura delas, que vai evoluindo. Nesse programa-jogo mostra-se como a moral do robô evolui. Temos, pois, que assumir que a programação da moral nas máquinas tem de vir a permitir a sua própria evolução[8]. Não há uma moral fixa, congelada. A moral é uma coisa evolutiva, e ao longo da história da espécie, tanto remota quanto próxima, tem vindo a desenvolver-se colectivamente.

É evidente que as máquinas estão cada vez mais autónomas, e temos que garantir que possam conviver connosco nos nossos termos e com as nossas regras. Há, portanto, um novo paradigma ético que diz que a moral também terá de ser computacional. Quer dizer, temos que ser capazes de programar a moral. Tal tem um lado positivo colateral, porque ao programarmos a moral nas máquinas percebemos melhor as nossas próprias éticas humanas.

Considere-se o seguinte exemplo, em trabalhos científicos sobre a culpa. Quando ela é introduzida em populações de agentes informáticos, estes passam a usufruir dessa capacidade. A sentirem-se coagidos quando fazem algo que prejudique outro elemento, donde resulta uma espécie de autopunição, uma mudança de comportamento, com o objectivo de evitar culpabilizações futuras. Não é uma culpa no sentido existencial, freudiano, mas num aspecto mais pragmático de não ficarem satisfeitos com o que fizeram ao prejudicar outrem.

Introduza-se uma dose de culpa – nem de mais nem de menos – em apenas alguns agentes, numa população deles a interagir dentro de um computador, num jogo evolutivo. Sem a existência desta componente de culpa, a maioria tenderá a jogar egoisticamente, cada um querendo ganhar mais do que os outros, não se conseguindo por isso chegar ao um nível em que todos pudessem ganhar ainda mais. Mas este resultado desejável já se torna possível havendo uma dose de culpabilização inicial, que modifique comportamentos e se espalhe como boa estratégia a toda a população. Mostrámos, matematicamente, que uma certa dose de componente de culpa é vantajosa, e promove a cooperação. Mas também que não se deve sentir culpa face a quem não se sente por sua vez culpado, pois tal é deixar-se sofrer abusos.[9]

---

[7] Pode ser visto neste link: https://drive.google.com/file/d/0B9QirqaWp7gPUXBpbmtDYzJpbTQ/view?usp=sharing, explicado em detalhe aqui: https://userweb.fct.unl.pt//~lmp/publications/online-papers/lp_app_mach_ethics.pdf.

[8] O robô vai mostrando num balão o que está a pensar, e mostra-se como o utilizador lhe vai dando novas regras morais a juntar às anteriores, por vezes suplantando-as quando existe contradição entre elas.

[9] Para os detalhes técnicos consulte-se: L. M. Pereira, T. Lenaerts, L. A. Martinez-Vaquero, T. A. Han, Social Manifestation of Guilt Leads to Stable Cooperation in Multi-Agent Systems, in: Procs. 16th Intl. Conf. on Autonomous Agents and

É este, aliás, o grande problema abstracto central da moral e do gregarismo, que naturalmente incide também no caso das máquinas: Como conseguir evitar o egoísmo puro dos agentes que oportunisticamente se querem aproveitar do gregarismo dos outros sem, por sua vez, contribuir para ele? Por outras palavras, como poderemos demonstrar, através de modelos matemáticos computacionais, em que circunstâncias o gregarismo é evolucionariamente possível, estável e vantajoso? E lograrmos usar o próprio computador para melhor entendermos como a maquinaria da culpa funciona, entre que valores de que parâmetros, e fazermos variar tais parâmetros para perceber como melhor os usar evolutivamente. Ao criarmos agentes artificiais que têm uma certa dose de culpa damos, ao mesmo tempo, argumentos ao facto de a culpa ser uma função útil, resultado da nossa evolução.

Como já se percebeu, estamos perante um tema que vive paredes meias com a Filosofia, com a Jurisprudência, com a Psicologia, com a Antropologia, com a Economia, etc., em que são importantes a interdisciplinaridade e a inspiração que nos dão esses vários domínios. É muito importante chamar a atenção para o facto de que um dos problemas que temos é o de a Jurisprudência não estar a avançar o suficiente, tendo em vista a urgência em se legislar sobre as máquinas morais. Quando se fizer legislação com respeito às máquinas, teremos que começar por definir e usar conceitos novos, sem os quais será impossível fazer as leis, pois estas têm sempre que apelar aos conceitos da Jurisprudência. E é relevante reconhecer que o Legislador está muito atrasado em acompanhar o passo da técnica. Tal é preocupante porque é vulgar uma confusão: a noção de que o progresso técnico é igual ao progresso social. Na verdade, o progresso técnico não está a ser acompanhado por um desejável e concomitante progresso social. A técnica deveria estar a ser usada ao serviço dos valores humanos, e esses valores devem ser desfrutados igualmente por todos, com a criação de riqueza a ser distribuída com justiça.

A História pode dar-nos grandes referências e linhas gerais de actuação. Por exemplo, se considerarmos o grande progresso e apogeu na civilização Grega, séculos V e IV a.C., constataremos que o mesmo só foi possível porque suportava-o uma legião de escravos, sem direitos de cidadania nem possibilidade de ascensão social, que eram constituídos pelos exércitos conquistados e pelos cidadãos estrangeiros. Ora, analogamente, temos a possibilidade de usufruir cada vez mais de máquinas escravas, que já o são, a desenvencilharem-nos do esforço que pode ser feito por elas. Mas gostaríamos que toda a gente fosse liberta e ganhasse igualmente com isso, através de uma distribuição justa da riqueza produzida por tais escravos, que – pelo menos por enquanto – não nos levantam qualquer problema de natureza ética. Ora, está a acontecer o contrário. As máquinas substituem as pessoas, resultando daí um lucro cada vez maior para os seus donos. A devida contraparte, ou seja, uma distribuição justa da riqueza adicional, está cada vez mais longe de acontecer. Ao passo que o universo de situações em que o humano não tem qualquer possibilidade de competir com a máquina não pára de aumentar. Daí ser indispensável o novo contrato social, em que a relação trabalho/capital seja reformulada e actualizada, em consequência do impacte social das novas tecnologias, nomeadamente do aumento de sofisticação das máquinas com cognição e autonomia.

Se uma máquina me vai substituir deverá fazê-lo completamente (até nas obrigações sociais). No sentido em que, Eu, ao posicionar-me numa actividade de trabalho, contribuo para a Segurança Social que sustenta os reformados actuais; contribuo para o Serviço Nacional de Saúde; contribuo com o IRS para tornar possível a governação e desenvolvimento do país, etc. Assim, se uma máquina me substituir completamente, eliminando-me de um trabalho cuja

---

Multiagent Systems (AAMAS 2017), Das, S. et al. (Eds.), pp. 1422-1430, 8–12 May 2017, São Paulo, Brazil. https://userweb.fct.unl.pt/~lmp/publications/online-papers/guiltEGT.pdf

actividade ela mantém, também deve pagar os impostos que eu estava pagando para sustentar o contrato social vigente. Substituir, tem de significar substituir nesses aspectos todos!

Não é possível reduzir os problemas da ética das máquinas a um código deontológico que os engenheiros informáticos devam seguir. Justamente pelo impacto que isso tem nos valores humanos e organização social, e no nosso devir civilizacional. Por isso, a questão dos valores é ineludível, e não se pode reduzir a meros 'standards' técnicos.

Há vários e repetidos relatórios de estudos, compagináveis, de entidades insuspeitas, nomeadamente da McKinsey & Company, do Pew Research Center, da OCDE, da PricewaterhouseCoopers, etc., que apontam para um acréscimo final de entre 15 e 20% de desemprego adicional em 2030, só em virtude da IA. O tópico do desemprego causado pela IA nas próprias superpotências de IA, e que noutras paragens será mais gravoso, é bem analisado no recente livro de Kai-Fu Lee.[10]

Se não tomarmos as atitudes adequadas presentemente, podemos imaginar as linhas gerais de um futuro que não será nada prometedor. Não podemos esquecer que neste momento há médicos a treinarem máquinas para lerem raios-X, interpretar análises, examinarem sintomas e por aí adiante. Por esse mundo fora, uma multidão de profissionais altamente capacitados, da medicina à economia ao direito, estão a passar conhecimento humano para máquinas que o saberão replicar e utilizar. As pessoas estão a ensinar quem as vais substituir.

Os perigos da IA não se configuram na possibilidade de aparecer um Exterminador. Os riscos consubstanciam-se no facto, de já neste momento, estarem máquinas simplistas a tomar decisões que nos afectam. No entanto, por lhes chamarmos "máquinas inteligentes" as pessoas acham que estão a fazer um bom trabalho. Este actual excesso de venda da IA é muito pernicioso. Além disso, a IA que está agora a ser vendida não chega a um décimo do que é, e poderá ser de facto a IA. A IA a sério ainda está por vir, e será muito mais sofisticada do que a generalidade dos programas actuais, ditos de *deep learning*. Estes são programas simplistas, e não se devia estar a dar tanto poder a máquinas tão simples. Mas como substituem humanos, como vão substituir radiologistas, condutores de automóveis, camiões, pessoas nos 'call-centers', pessoas na segurança dos centros comerciais, são vendidos como uma panaceia.

O autor faz parte do projecto *Incentives for safety agreement compliance in AI Race* [11] patrocinado pelo *Future of Life Institute*[12], uma organização sem fins lucrativos. O projecto, na área da segurança de 'software', aborda a questão da urgência em chegar ao mercado, por parte das firmas que desenvolvem produtos de IA. Mais concretamente, analisa as consequências de se descurarem as condições de segurança desses produtos. A urgência é tal que a segurança é posta de lado, porque custa dinheiro e tempo, e atrasa a chegada ao mercado antes dos competidores. O objectivo do projecto relaciona-se com o estabelecimento de regras do jogo de forma a que não haja ninguém a fazer vista grossa em relação à segurança, como se ela não fosse essencial[13]. Para isso precisamos de entidades de regulação e monitorização, bem como a necessidade de uma "Comissão Nacional de Ética para a IA," que inclua a Robótica, por analogia a "Comissão Nacional de Bioética". Não podemos aceitar, como se ouve dizer na Europa e nos EUA, que as firmas que fazem carros sem condutor é que são exclusivamente responsáveis, e que se houver algum problema logo se verá. Os Governos não seriam, portanto, responsáveis pelos testes a que tais carros devam ser submetidos, simplesmente delegam nas

---

[10] Kai-Fu Lee, *AI super-powers – China, Silicon Valley, and the New World Order*, NY: Houghton Mifflin Harcourt, 2018.
[11] https://drive.google.com/open?id=1j59rhP7op3nBpvaxpeCdaBVObJAbzWBJ
[12] https://futureoflife.org
[13] Para resultados do projecto veja-se: T. A. Han, L. M. Pereira, T. Lenaerts, *Modelling and Influencing the AI Bidding War: A Research Agenda*, nas actas de: AAAAI/ACM Conference on AI, Ethics, and Society, (AIES 2019), 27-28 January 2019, Honolulu, Hawaii, USA. Aqui: https://userweb.fct.unl.pt/~lmp/publications/online-papers/AI race modelling.pdf

próprias empresas. Mas atente-se aos recentes acidentes ocorridos com o Boeing 737-Max [14], em que a *Federal Aviation Authority* (FAA) americana delegou na própria Boeing as verificações de qualidade!

Na União Europeia a responsabilidade pela segurança aparenta ser mais disfarçada. Criou-se uma comissão de alto nível para a IA e Ética para dar recomendações[15]. O que se propõem em súmula produzir é: "Temos aqui umas recomendações que as firmas de desenvolvimento poderão seguir. Teremos além disso firmas credenciadas privadas de auditoria que vão inspecionar aquelas firmas." Iremos, porventura, cair no esquema de auditores com interesses nas próprias entidades que examinam, porque estas também lhes encomendam estudos. Veja-se o caso dos bancos e da crise financeira de 2008.

Não fecharemos este texto sem uma pequena síntese dos "cadernos de encargos" lançados à comunidade científica da IA, e não só:

- Precisamos saber mais sobre as nossas próprias facetas morais para conseguirmos passá-las às máquinas. Contudo não sabemos ainda o suficiente sobre a moralidade humana. Nesse sentido, é importante reforçar o estudo dela pelas Humanidades e Ciências Sociais.
- A moralidade não é apenas acerca de evitar o mal, mas também acerca de como produzir o bem. Maior bem para maior número de pessoas. A problemática do desemprego é inerente a este ponto de vista.
- As universidades são o sítio apropriado para abordar todas estas questões, pelo seu espírito de independência, a sua prática de raciocínio e discussão. E contêm, nas suas faculdades, a necessária interdisciplinaridade.
- Tão cedo não vamos ter máquinas com uma capacidade moral geral. Teremos máquinas que sabem respeitar normas num hospital, numa prisão, e até as normas de guerra. Estas são até as mais bem particularizadas e subscritas por todo o mundo. Como estão bem especificadas são menos ambíguas e estão mais próximo de poderem ser programadas.
- Iremos começar por automatizar as normas e as suas excepções, pouco a pouco, alargando a generalidade e a capacidade de uma máquina para aprender novas normas, e de ampliar, evoluindo, os seus domínios de competência, com a imprescindível segurança.

Do ponto de vista dos critérios de acção, a moral alcandorada nos céus do passado está confrontada com uma nova perspectiva sobre os sistemas morais nascentes, estudados no âmbito da psicologia evolucionária e aprofundados através de modelos testáveis em cenários artificiais, como agora permitido pelos computadores. À medida que a investigação avance, podemos conhecer melhor os processos inerentes à decisão moral, a ponto de eles poderem ser "ensinados" a máquinas autónomas, capacitadas para manifestarem discernimento ético.

No domínio da Economia há toda uma problemática associada ao impacte no trabalho e à dignidade que lhe é inerente, bem assim como à produção e distribuição da riqueza; ou seja, toda uma reconfiguração das relações económicas que resultará, não apenas da automação de actividades rotineiras, mas fundamentalmente da entrada em cena de robôs e 'software' que poderão substituir médicos, professores, ou assistentes em lares de terceira-idade (para darmos nota de profissões as quais o olhar comum não percepciona como facilmente supríveis). O conhecimento deste âmbito é especialmente relevante, exigindo tomadas de posição que sustentarão a necessidade de uma moral social actualizada, e um renovado contrato social.

A problemática da moral computacional ganha assim existência num contexto em que o ecossistema do conhecimento estará bastante enriquecido, pois terá de incorporar agentes

[14] Para poupar custos. Ver "*Boeing's 737-Max software outsourced to Rs 620-an-hour engineers*". Aqui: https://economictimes.indiatimes.com/industry/transportation/airlines-/-aviation/boeings-737-max-software-outsourced-to-rs-620-an-hour-indian-engineers/articleshow/69999513.cms?from=mdr

[15] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

não-biológicos com capacidade para se tornarem intervenientes activos em dimensões que, até agora, têm sido atribuídas exclusivamente a humanos.

Sendo assuntos muito difíceis, quanto mais cedo começarmos melhor!

Agradecimentos

Este texto adveio da iniciativa *Diálogos Intergeracionais* initiative: https://www.cnc.pt/dialogos-intergeracionais-mesa-redonda-4/ e a ser publicado no e-livro resultante.
O autor agradece o apoio do projecto RFP2-154 do "Future of Life Institute", USA; e do projecto FCT/MEC NOVA LINCS PEst UID/CEC/04516/2019 da "Fundação para a Ciência e a Tecnologia", Portugal.