Cavalgando o "Big Data"

Poderia ser o título de mais um Western, glorificando os cowboys civilizadores e demonizando os atrasados "índios" pouco entendidos em novas tecnologias desde as "Colt" e "Winchester" até aos líquidos mais ou menos destilados.

Mas não. Vem este pequeno texto a propósito do filme "The Great Hack" onde o papel da empresa Cambridge Analytica (CA), uma então florescente companhia que usava enormes quantidades de dados sobre indivíduos para tentar influenciar eleições, foi desmascarada e acusada de um dos maiores escândalos socio políticos dos últimos anos.

Não sendo especialista em estética cinematográfica, abordarei mais o tema ilustrado que o filme como produto aliás notável e muito eficiente.

A ideia mais relevante que emerge deste filme é mais uma evidência de como a ganância de alguns, no filme ilustrados pelo ex CEO da CA Alexander Nix, a frustração ressentida, transformada em arrependimento (honesto ou exibicionista) de outros, aqui representados por Brittany Kaiser e Chris Willie, e a ingenuidade acrítica de muitos, aqui representados por votantes anónimos, podem conjuntamente concorrer para se chegar a resultados socialmente catastróficos. E foi esse o caso do resultado das eleições presidenciais americanas de 2016.

Este filme relata uma realidade onde fica evidente o "dark side" da má utilização de uma combinação de metodologias, desde a nova Tecnologia Computacional envolvendo subáreas da Inteligência Artificial (IA) baseadas em "Big Data", certas vertentes da Psicologia, no filme definida como arma já anteriormente muito usada, que

transitou do cenário de guerra para o das Eleições, e ainda a "Ciência" Política, desenvolvidas todas com o bem em mente, mas aqui usadas com o objetivo primeiro de fazer dinheiro e conquistar poder a todo o custo.

O uso de dados pré-existentes para classificar e prever é uma realidade com tradição de permitir alcançar bons e fiáveis resultados. Há mais de 30 anos que, eu próprio, utilizei grandes quantidades de dados, vindos até dos USA para, por exemplo, prever futuros consumos de água numa determinada cidade ou região. Tal previsão permitia saber de antemão o que a municipalidade deveria bombear e armazenar para ser consumido em cada semana, garantindo assim um serviço vital às populações sem desperdícios inúteis. Ou ainda para, a partir dos dados obtidos com muitos tipos de manchas na pele, ajudar a diagnosticar aquelas que poderiam tornar-se cancerígenas. A tecnologia envolvida em tais sistemas aplicados já incluía processos e algoritmos com os mesmos fundamentos científicos da atual tecnologia, agora em causa.

Não há dúvidas que o filme é muito assertivo sobre os perigos da má utilização, com fins pelo menos discutíveis, de certas tecnologias que se tem revelado extremamente poderosas.

Mas o mundo nunca foi, e muito menos no presente, descritível só a preto e branco. Há quem use para o mal o que pode ser usado para o bem. E, no lado negro, torna-se evidente que as Cambridge Analytica, que por aí pululam, são os gangsters do mundo atual, jogando com as referidas frustrações, ganância e ingenuidades de tantos. Também Facebook e Google são, pelo menos, os cúmplices e facilitadores do "grande negócio". E só há uma atitude cívica válida: Há que os confrontar, defrontar e vencer. Mas, nesta luta, há riscos evidentes. E o menor não será o de poder repetir os erros daqueles grupos de

trabalhadores desesperados (os Luditas) que no início da Revolução Industrial, no sec. XIX, pensavam resolver os seus problemas, verdadeiros e dolorosos, advogando a destruição das máquinas, demonizando a tecnologia que os substituía nos seus postos de trabalho.

Mas como é que a situação atual torna tudo isto, poderes e perigos, possível?

Precisemos que, quando falamos em Big Data, estamos a referir informação representada e medida em quantidade de bits da ordem dos petabytes, ou seja de 10 elevado à potencia 15. Também poderemos dizer que é cerca de 1 Milhão de Giga Bites. Portanto uma coleção enorme de dados.

E quando hoje se fala em dados não pensemos que estamos a referir informação apenas contida em Bases de Dados com tabelas de informação estruturada. Não. Hoje, qualquer texto corrido, qualquer segmento de voz, discurso ou imagem, desde que em formato eletrónico, são passíveis de ser considerados dados e de serem tratados, analisados, manipulados, correlacionados, contribuindo assim com informação para um conhecimento mais abrangente.

E se a posse de tantos dados, juntamente com os algoritmos, mais ou menos públicos que os tratam, são o novo petróleo, provavelmente serão as empresas maiores (com mais clientes ou utentes, como as Amazon, Facebook, Google) e também os países mais populosos (como China, Índia, USA, Rússia, Japão) que terão sempre maior quantidade de dados à disposição e, em consequência disporão, no futuro, de maior poder face a todos os outros.

Não duvidem que os Algoritmos de tratamento dos Dados são, não direi perfeitos, mas muito corretos e confiáveis, no sentido que inclusivamente medem a probabilidade de erro associada a cada decisão e classificação. Portanto devemos considerar tais algoritmos como produtos científicos válidos e de grande alcance aplicacional.

É verdade que nos últimos anos, menos de 10, se deu uma grande alteração na escala de possibilidades da tomada de decisões automatizadas devido a 3 fatores essenciais:

- A existência de cada vez maior quantidade de dados produzida por todos nós e coligida por alguns; o que é óbvio no filme.

- a cada vez maior capacidade de integração dos circuitos computacionais (o hardware) o que aumenta desmesuradamente a sua potencia. Não nos esqueçamos que o relatório oficial dos USA sobre os acontecimentos do "11 de Setembro" concluiu da prévia existência de informação e dados suficientes para ter sido possível prever o que aconteceu, ao que as agências NSA, FBI, CIA responderam ter sido impossível ter detetado tal hipótese em tempo útil. Pois bem, hoje essa desculpa já não seria válida dado o poder computacional que tais agencias têm instalado.

- e finalmente o desenvolvimento em IA de Algoritmos mais capazes da propriedade de aprendizagem a partir de dados. Algoritmos baseados em Redes Neuronais Artificiais, agora muito modificados e a que chamamos "Deep Learning" e que não degradam o desempenho com o aumento da quantidade de dados, são as estrelas atuais.

Tais algoritmos, também referidos como de "Machine Learning" ou ainda de "Data Mining" que, no essencial, são sub-áreas da IA, nas suas várias dimensões e paradigmas, também devem ser considerados

muito relevantes pelo conhecimento útil que podem extrair de dados cuidadosa e legalmente colecionados.

Decisões de Empresas que assim podem afinar os seus produtos para melhor satisfação do cliente (e isso pode ser bom ou não, depende dos objetivos dos produtos e dos dados utilizados) ou, por exemplo, diagnósticos médicos que passam a ser muito mais precisos, são ganhos civilizacionais apreciáveis. Epidemiologistas podem retirar padrões e correlações entre fármacos, as vacinas administradas e as condições específicas dos pacientes assim projetando tendências e propondo ações apropriadas. Foi o que a EMA fez relativamente a determinar os efeitos das vacinas contra a Covid19. Mas também assuntos mais prosaicos, como a gestão de trânsito em grandes áreas metropolitanas que pode melhorar muitíssimo com tais análises. A lista das aplicações benéficas é enorme.

A outro nível, governos e grandes companhias já exploram este "Big Data" em seu proveito, mas anunciando que é também em nosso proveito. Melhores produtos, luta contra o terrorismo, previsão de alterações climáticas, etc.

É claro que se formos nós próprios a produzir e voluntariamente enviar dados constantemente, por telemedicina, por exemplo de um pacemaker que está sempre a recolher e enviar os nossos sinais biométricos para um médico ou clínica e assim poder receber avisos que serão úteis na prevenção de acidentes cardíacos emergentes, vemos isso, indiscutivelmente, como um progresso. Não nos importaremos de enviar esses nossos dados. Poderão dizer que são só uns poucos dados pessoalmente voluntariados. Mas atenção, pois para que um programa computacional possa reagir a um acontecimento simples e pessoal, como uma determinada onda cujas características antecipam um determinado evento, e assim fazer uma previsão

acertada, terá sido preciso, anteriormente, ter analisado um histórico de milhares de casos. Dados de outros …

Há até quem garanta, e o escreva, (ver "The Master Algorithm", P. Domingos da Universidade de Washington) que a cura definitiva do cancro em geral vai ser conseguida pela correlação de toda informação desde já existente em artigos científicos (palavras, logo dados) em conjunto com os resultados laboratoriais já recolhidos, e sua correlação com os princípios ativos de fármacos já desenvolvidos. Os algoritmos irão chegar lá. Talvez…

Modelos novos de negócio, como os da Netflix e da Uber, que apostam no exame de dados dos consumidores para melhor lhes responderem, podem ser apresentados como sendo de sucesso. O outro lado da moeda é que as anteriores intervenientes no negócio (aluguer de vídeos ou taxistas tradicionais) tenderão a estiolar. Podemos ver estas alterações como uma perda ou como uma renovação constante, agora acelerada, da sociedade.

Portanto, com maior ou menor exagero e otimismo, também há muito boas notícias que podem vir desta tecnologia.

No entanto,

muitos consideram que os algoritmos que conseguem aproveitar tal massa de informação que o "Big Data" torna disponível são, sobretudo, potencialmente muito perigosos.

Uma vez que todos estamos, e a toda a hora, a produzir eventos com rasto eletrónico, ao telemóvel, ao computador, na nossa movimentação e localização, nas compras, na passagem em ruas e praças das cidades equipadas com câmaras de vídeo, no emprego, na escola, nas lojas, na clínica,… todos esses dados podem ser coligidos,

colecionados e correlacionados. E possivelmente não o deviam ser assim indiscriminadamente.

Utilizadores, compradores, empregados, simples cidadãos enquanto tal, estão a ser classificados de acordo com certos grupos comportamentais (que é um dos resultados desses algoritmos) considerados mais ou menos desejáveis ou indesejáveis. Aí perguntamos: E a que propósito? Ergam-se com a divisa: "Abaixo a exploração encapotada dos nossos dados"!!

Sim, mas até estávamos avisados. É que nada disto é novo para quem se habituou há muitas décadas a ler os autores mais ou menos futuristas, sejam o George Orwell ou o Aldous Huxley, ou todos aqueles escritores e realizadores de ficção científica que me abstenho de nomear. Estava anunciado!

O que se passa agora é que juntando às potencialidades técnicas, a ingenuidade de muitos, que, por exemplo, respondem diretamente a inquéritos para traçar o seu perfil, é claro que se dá aso, como se comprova pelo filme, à manipulação comportamental, permitindo Interferências em eleições, detetando primeiro e Influenciando depois quem é mais suscetível de ser influenciado (os ditos "persuadables").

Novos votantes por exemplo. Se enviam informação que facilmente se pode traduzir em fatores psicológicos para a construção do perfil ou personalidade indutora dos comportamentos, então tanto permitem a felicidade dos bruxos de esquina que enganam os patego, como a dos mais ou menos sofisticados influenciadores de opinião. Possuindo essa informação, o que se segue será a geração de "Fake News" à medida dos desejos adivinhados ou explícitos daqueles grupos de votantes, acertando no "turning point" que lhes vai virar a opinião (alterar a polaridade, diz-se).

E tais procedimentos são, no meu entender, claramente criminosos. Mas deturpar notícias com arte para influenciar o cidadão é o que mais temos visto e sofrido nos "media", e haverá sempre advogados a provar que não será bem assim, que são apenas diversas perspetivas "criativas", que há sempre um fundo de verdade naquelas notícias, que estão a exercer a liberdade (mesmo de enganar) e os outros é que são manipuladores, etc.

Dados, classificações, perfis, notícias (falsas ou não) à medida, foi precisamente o que fez a Cambridge Analytica, assumindo-se nitidamente como uma empresa de propaganda política. Enviou uns milhares (não milhões) de Inquéritos para construir perfis, aliás sem grande imaginação, pois há quem o faça bem melhor, em formato de jogos interativos e mais atrativos e, mais tarde, pôde enviar mensagens apropriadas para cada grupo de votantes diferentemente classificado. E deu, como sempre dá, resultado em, pelo menos, alguns dos alvos!

Não nos esqueçamos que campanhas dos noticiários nas TVs o fazem diariamente, em maior ou menor grau, para os seus espectadores. Selecionam, alinham e deturpam, tentado encontrar o ângulo que fará mudar a opinião dos seus espectadores, no sentido previamente especificado e desejado.

Alarmante? Mas nós não nos importamos de, ao longo da nossa vida, ter contribuído para dossiês médicos sobre nós próprios ou mesmo sobre os nossos filhos. Pensamos sempre que é para o bem. Também temos associados a cada um de nós, grandes, e muitas vezes longos, registos escolares, cheios de classificações, de pontuações que são muitas vezes usados, quer para nos dar, quer para nos retirar, oportunidades. E todos achamos isso aceitável. Ou seja, nós já somos pontuados ao longo da vida de acordo com um histórico de eventos, e as nossa possibilidades, nos futuros empregos e até muitas vezes na

vida privada, dependem em muito desses percursos, destas classificações, enfim destes perfis.

Não recuemos, portanto, com horrores de virgens quando ouvimos falar de pontuar o desempenho das pessoas (os "scores" de crédito social). Claro que tudo depende do sistema instituído para isso, sua transparência, as nossas defesas disponíveis, possibilidade de crítica e base científica para a sua aceitação ou não.

Muito importantes são também as questões de quem serão os proprietários dessas grandes quantidades de dados, quem pode aceder a tais informações e para que fins poderão ser usadas.

Há uma dimensão de risco diferente se os dados são pessoais e identificáveis ou se são anonimizados e tratados apenas para retirar tendências genéricas de uma população de clientes, de votantes, de pacientes, etc.

Uma medida fundamental a garantir seria a de que cada um ser o único dono dos seus dados pessoais e, fora isso, só dados completamente anonimizados deveriam poder ser guardados e utilizados para extração de padrões, tendências, classes, correlações Ou seja, a PRIVACIDADE, deverá ser ou equivalente a impedimento de coligir dados, ou deveria ser equivalente à garantia exclusiva de propriedade por cada um (pessoa, instituição, empresa …) dos seus próprios dados.

Como sabemos, já está a acontecer que entidades detentoras de coleções de dados, para além de as usarem para tomar decisões proveitosas para si, as vendem a terceiros para serem usadas como bem o quiserem. E, em face desta realidade, há quem se preocupe não com o facto em si e suas consequêncuas, mas, acreditem, apenas em partilhar o lucro daí resultante …

Não cabe aqui a problemática de como Companhias de Seguros estão ávidas de conhecer não só o nosso historial de doenças mas também os nossos hábitos, seja de mobilidade, de condução, desportivos, entre outros, para taxar os seus produtos em conformidade com o risco avaliado por dedução para cada um. Conduz de noite ou em estradas difíceis? Não será grande cliente! Ou seja, no limite só aceitariam ter lucro com cada um dos clientes, impedindo qualquer risco da parte da Companhia.

Ou ainda de como os Bancos tentam usar a maior quantidade de dados possível sobre os clientes para tomarem as suas decisões. E, nesse caso haverá decisões aceitáveis e outras que o são menos. Por exemplo, já aconteceu que cartões de crédito baixaram automaticamente o seu teto de crédito, não por um cliente ter uma história pessoal de grande endividamento, mas por usar esse cartão em lojas onde alguns outros já se endividaram demasiado. Dedução automática abusiva.

Todos devem saber que a NSA dos USA, mas também os seus equivalentes russos, chineses e israelitas, e sabemos lá quantos mais, usam os nossos dados e os dos nossos contactos em redes sociais, os contactos no Linkedin, ou os "amigos" no Facebook, para nos espiar e classificar. Somos ou não um POTENCIAL Terrorista? Claro que justificando que estão a evitar que algum terrorista embarque num avião em que poderemos também viajar, e assim nos proteger a todos, já aceitamos melhor tal devassa …

Finalmente, toda esta informação está localizada não já em computadores específicos e pessoais ou mesmo das empresas que as usam, mas na Núvem (na "Cloud") podendo ser mais ou menos facilmente acedida, desviada e apropriada por Hackers.

Pior: E se esses Hackers modificarem alguns aspetos cruciais de certos dados de forma a influenciar as conclusões que o algoritmo retira sobre nós de uma maneira impecavelmente correta do ponto de vista científico?

Chama-se a isso "Counter AI" e é um desporto de algumas Agências governamentais por esse mundo. Como se tornará fácil exclamar: "Eis a prova de que aquele país possui armas de destruição massiva" … que conclusão conveniente …

Daí que, tão importante como ter algoritmos transparentes e corretos que nos permitam melhorar a vida em sociedade, e temos grandes esperanças no futuro da IA, é fundamental velar pela integridade dos dados, caso eles devam mesmo existir e ser conservados, o que será difícil de impedir pelo menos em certa medida. Os chamados Curadores da veracidade dos dados são de suma importância na garantia da integridade dos cidadãos e dos factos.

Em conclusão, é fundamental afirmar que todos estes perigos implicam a urgente necessidade de uma maior consciência cívica do cidadão em geral, assim como dos cientistas também, tornando necessário uma mais apropriada legislação restritiva das operações de coleção, conservação, utilização e transação dos nossos dados pessoais. Proteção sobretudo das nossas interações, controlando e monitorando os óbvios perigos inerentes à sua utilização abusiva sem, ao mesmo tempo, impedir os fins científicos que sejam aceitáveis para a humanidade.

Por isso, e para terminar com polémica, diria que este filme, "The Great Hack", é uma peça importante de denúncia de grandes perigos atuais e futuros, sem deixar de ser um tanto simplista na demonização da tecnologia usada para fins criminosos.

Eugénio Oliveira

Prof. Catedrático Emérito da Universidade do Porto