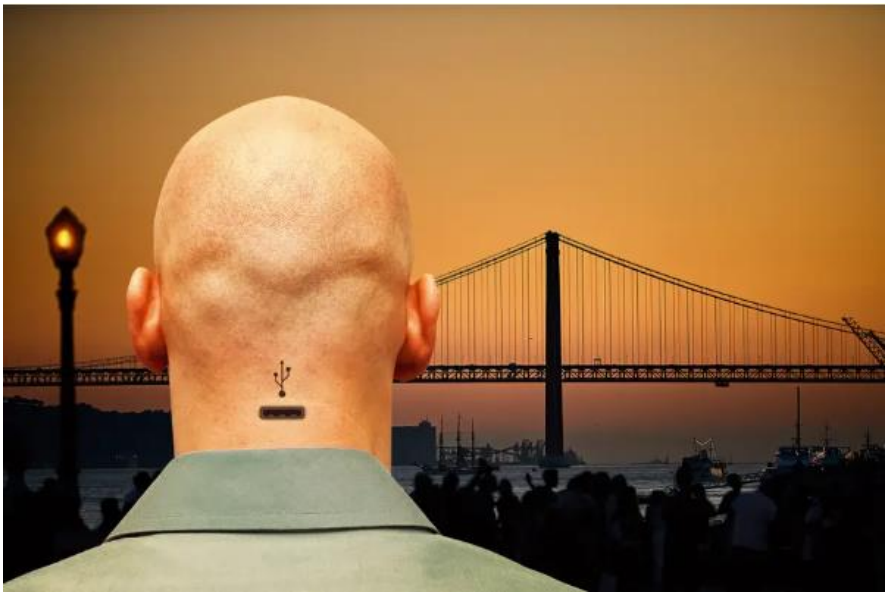

O maior risco da inteligência artificial é não haver suficiente

Autor: Pedro Domingos

24 de Junho de 2023

A inteligência artificial (IA) arrisca-se a conduzir à extinção da Humanidade? Há quem diga que sim, incluindo nomes famosos da tecnologia como Elon Musk e pioneiros da IA como Yoshua Bengio. Segundo esta teoria, se criarmos máquinas mais inteligentes do que nós estas poderão escapar ao nosso controlo e decidir exterminar-nos. E desenvolvimentos recentes, como o do ChatGPT, convenceram alguns de que este não é um perigo remoto, mas algo que se aproxima rapidamente e requer intervenção urgente dos Governos. Antes de mais, dizem, é preciso limitar e retardar o progresso da IA, sob pena de perdermos o controlo do planeta.

Antes de cairmos na histeria, no entanto, seria bom examinar com algum cuidado os pressupostos subjacentes a este cenário apocalíptico. Um erro bem conhecido na IA é a chamada “falácia do homúnculo”, a ideia de que dentro de cada sistema de IA existe um “mini-homem” com todas as características dos seres humanos: emoções, desejos, preconceitos,

intenções, motivações, autonomia, até mesmo a consciência. Este erro é natural e muito comum, especialmente entre os não especialistas de IA, porque os seres humanos raciocinam por analogia e quando veem um sistema comportar-se de forma inteligente imediatamente projetam nele as outras características das únicas entidades inteligentes que conhecem: os seres humanos e outros animais. O erro é agravado por Hollywood, em cujos filmes as IA e os robôs são sistematicamente seres humanos mal disfarçados, porque essa é a forma mais fácil de fazer filmes que sejam interessantes para nós.

É no entanto um erro completo. A inteligência artificial é extremamente diferente da humana, e sem compreender isto é difícil compreender quais são ou não os seus riscos. Os algoritmos de IA são apenas isso: algoritmos, ou seja, sequências de instruções dadas por nós que o computador executa sem se desviar um milímetro. O ChatGPT, por exemplo, não faz mais do que prever a próxima palavra num texto tendo em conta as palavras anteriores, e assim, palavra a palavra, cria um novo texto a partir do que lhe é dado.

O alarmismo em torno da IA foca-se em particular nos algoritmos de aprendizagem automática (a minha especialidade desde há 30 anos). Na aprendizagem automática, o computador desenvolve os seus próprios algoritmos a partir dos dados. Pode, por exemplo, aprender a diagnosticar o cancro do pulmão a partir de radiografias e dos correspondentes diagnósticos produzidos por médicos, ou, como o ChatGPT, a preencher os espaços em branco num texto a partir de outros textos. A fonte do alarme é a ideia de que quando o computador começa a aprender por si próprio é impossível prever o que vai aprender e pode começar a fazer coisas arbitrariamente malignas. Mas isto é uma ilusão. Os sistemas de aprendizagem automática, como muitos outros, são governados por objetivos fixos como maximizar o engagement (no caso das redes sociais) ou minimizar a taxa de erro nas previsões (no caso do ChatGPT). O sistema faz uma vasta procura para encontrar o melhor modelo de acordo com estes critérios e a cada passo os modelos que não constituem melhorias são deitados fora. É física e matematicamente impossível que o sistema decida por si próprio começar a evoluir noutras direções, por exemplo para satisfazer os seus próprios desejos (que não tem). A aprendizagem automática é semelhante à reprodução selecionada, exceto que em vez de produzir plantas ou animais domesticados produz algoritmos. Ninguém teme ser morto uma noite pelo seu cão, mas os cães são apenas lobos domesticados. Porquê então temer ser morto por um robô cujo processo de evolução foi muito mais completamente controlado por nós?

Mas, insistem os alarmistas, e se a IA decidir que a melhor forma de atingir os seus objetivos é exterminar a Humanidade? Mas o problema aqui não é a maldade da IA, mas a sua

incompetência. E a solução para a incompetência é mais inteligência, não menos. E se, apesar de tudo, a IA seguir o caminho errado? Sendo muito mais inteligente do que nós, não nos será impossível detetá-lo? Não. Tecnicamente, a IA é o subcampo da informática que procura resolver problemas intratáveis, isto é, problemas que, no pior caso, demoram tempo exponencial a resolver, mas cujas soluções podem ser rapidamente verificadas.

Ou seja, mesmo que as IA sejam exponencialmente mais poderosas do que nós, continua a ser fácil verificar que não se afastaram do bom caminho. Uma das grandes potenciais aplicações da IA é curar o cancro. Quanto mais inteligente um sistema para este efeito for, melhor. Queremos mesmo limitá-lo na base de que magicamente decidirá ao invés começar a matar os pacientes e, mais absurdamente ainda, que nós não consigamos detetar isso?

Ainda mais risível é a noção que a tomada de poder por IA rebeldes é um perigo iminente. A inteligência real do ChatGPT é muito menor do que a aparente, e está ainda muito longe da humana; os resultados surpreendentemente bons que frequentemente produz derivam principalmente de ter processado uma quantidade inimaginável de dados. Desde os princípios da IA nos anos 50 tem havido sempre uma marcada tendência, mesmo entre os especialistas, para subestimar pronunciadamente a dificuldade do problema; a inteligência parece fácil porque é em grande parte subconsciente, mas evoluí-la levou biliões de anos e o cérebro humano é a entidade mais complexa do universo. E mesmo que uma malévola IA ao nível da humana surgisse amanhã, teria poucas hipóteses contra nós. Os nossos cérebros têm maior capacidade computacional total, as nossas necessidades de energia são vastamente menores, a nossa mobilidade e dexteridade são muitíssimo melhores, aprendemos mais a partir de muito menos dados, temos milhões de anos de experiência de combater adversários, etc. A possibilidade de a IA exterminar a Humanidade pertence aos filmes de ficção científica, não à realidade.

A IA E A DESINFORMAÇÃO

Os riscos inerentes à IA não se limitam aos existenciais. Outro que tem recebido uma quantidade desproporcionada de atenção é o de que a IA conduzirá a um aumento maciço da desinformação online, os humanos serão facilmente manipulados por ela, e isto será o fim da democracia. A realidade comprovada por vários estudos, no entanto, é que a desinformação online pouca diferença faz. Os mass media, por exemplo, influenciam muito mais os eleitores. Além disso, a quantidade de desinformação online é já vasta, e o limite principal é o da nossa atenção, não a capacidade de produção que a IA aumentaria. Mas, dizem os alarmistas, a IA será capaz de manipular os humanos muito melhor, e aí é que está o perigo. Felizmente, os humanos são

menos estúpidos do que os bem-pensantes pensam. Nos anos 50 a preocupação era que os anúncios de televisão nos manipulariam com grande facilidade, mas as pessoas rapidamente aprenderam a não lhes dar credibilidade, e o mesmo começa já hoje a acontecer com o ChatGPT. Todos sabem que ele frequentemente diz falsidades, e manipulá-lo tornou-se um passatempo com resultados frequentemente hilariantes. Pelo menos para já, os humanos são muito melhores a manipular as IA do que o contrário.

Mais importante, no entanto, é que a melhor forma de combater a desinformação online é precisamente através da IA. A ideia de que a IA gerará desinformação em grande escala é uma hipótese sobre o futuro; o papel essencial da IA na deteção de desinformação online é uma realidade hoje. O grande problema é que em muitos casos a IA ainda não é suficientemente sofisticada para detetar a desinformação, e esta passa pelos filtros automáticos em quantidades excessivas para ser detetada por operadores humanos. A solução, mais uma vez, é mais inteligência, não menos.

Outro hipotético grande risco da IA é o de "perpetuar os enviesamentos". A IA é supostamente um instrumento de opressão e discriminação, e um obstáculo à justiça e à equidade. Todos os dias parecem trazer novos exemplos de sistemas de IA tratarem homens e mulheres, ou brancos e negros, de forma diferente. Os denunciantes destes exemplos, no entanto, sistematicamente confundem a correlação com a causalidade, atribuindo a enviesamentos toda e qualquer diferença de resultados. Estas diferenças têm tipicamente boas razões. Por exemplo, refletem a realidade: se há mais engenheiros do sexo masculino, queremos que os sistemas de IA nos comuniquem esse facto ou que construam um alternativo mundo orwelliano em que o número de engenheiros dos dois sexos é igual? Ou as diferenças refletem propriedades da aprendizagem automática que nada têm a ver com os enviesamentos. Por exemplo, se 90% da população são de uma raça e 10% doutra, é óbvio que o modelo aprendido para a primeira será provavelmente melhor do que o aprendido para a segunda. Discriminar contra uma para compensar a outra, como alguns defendem, seria perverso; o objetivo deve ser aprender o melhor modelo possível de cada indivíduo, seja qual for o seu sexo ou a sua raça.

A realidade é que a IA é uma oportunidade incomparável para combater os enviesamentos e a discriminação. Os algoritmos de aprendizagem são matematicamente incapazes de ter enviesamentos relacionados com o sexo ou a raça, porque nem sequer sabem que estes existem. (Os dados podem ter enviesamentos, mas a solução aqui é obter amostras fiéis da população, um problema bem conhecido em estatística; e utilizar como dados valores reais, não

valores formulados hipoteticamente por humanos.) Em “Pensar, Depressa e Devagar”, Daniel Kahneman dedica um capítulo inteiro ao facto que os algoritmos tomam melhores decisões que os peritos humanos, e cita: “Não há nenhuma controvérsia em ciências sociais em que uma quantidade tão grande de estudos qualitativamente diversos dá resultados tão uniformemente na mesma direção que esta.”

A IA E APRIVACIDADE

Outro grande mal da IA é supostamente a invasão da privacidade. Utilizando algoritmos de aprendizagem, as grandes empresas de tecnologia como a Google e a Meta inferem os nossos interesses a partir do nosso percurso online, e servem-se disso para escolher os anúncios que nos mostram. Isto é alegadamente uma exploração imoral dos utilizadores. “Se o produto é grátis, és tu o produto”, diz o ditado. A União Europeia ataca estas e outras utilizações dos dados com o Regulamento Geral de Proteção dos Dados (RGPD), uma lei que considera orgulhosamente um modelo para o mundo inteiro seguir. Mas esta ideia que a publicidade direcionada, como é conhecida, é nociva para os utilizadores combina a incompreensão da economia com a da tecnologia e mesmo a da psicologia. A publicidade direcionada não é um jogo de soma zero; o facto de, por exemplo, a Google não nos cobrar nada pela pesquisa na Web não implica que ela tenha de extrair o dinheiro de nós de outra forma. Pelo contrário, todos ganham com a publicidade direcionada: a Google, que é paga pelos anunciantes; os anunciantes, que ganham quando os utilizadores compram os seus produtos; os produtores de conteúdos, que são indiretamente pagos pelos anunciantes através da Google; e os utilizadores, que usufruem de graça de um serviço valioso (a pesquisa). E de todos estes os maiores beneficiários são de longe os utilizadores. O rendimento da Google por utilizador anda na casa das centenas de dólares por ano. O valor da pesquisa para o utilizador mediano, de acordo com um estudo do MIT, é da ordem das dezenas de milhares de dólares por ano — cem vezes maior.

Além disso, eu, como utilizador, prefiro grandemente ver anúncios relevantes a perder tempo com anúncios genéricos que não me dizem nada. O maior problema de sistemas como o da Google é que a sua capacidade de prever quais os anúncios que me vão interessar é ainda muito limitada, em parte porque os dados que a Google tem sobre mim são limitados, e em parte porque os algoritmos de aprendizagem podiam ser muito melhores. Mais uma vez, o problema é a falta de inteligência, não o excesso. E a nossa preferência como utilizadores deve ser fornecer à Google mais e melhores dados sobre nós, não “proteger a nossa privacidade”. Esta ideia que temos de proteger os nossos dados das malévolas empresas é semelhante à ideia que temos de

guardar o nosso dinheiro debaixo do colchão para que os bancos não o roubem. Os nossos dados devem ser investidos para nosso benefício, tal como o nosso dinheiro. A preocupação com a privacidade que as velhas gerações têm não é partilhada pelas novas, e a longo prazo a ideia de esconder os nossos dados das empresas parecerá tão ridícula como a de guardar o dinheiro debaixo do colchão. Mas o RGPD, longe de compreender isto, desperdiça o tempo dos utilizadores a autorizar cookies e mais cookies, sem quaisquer benefícios visíveis. Pior ainda, cria custos e entraves desnecessários às empresas, e em particular às startups, que têm menos capacidade económica. A preocupação principal da UE devia ser fomentar estas empresas e trazer para a Europa os enormes ganhos económicos da tecnologia avançada, mas pelo contrário, não pára de lhes dificultar a vida. Uma das principais imposições do RGPD é que os dados só podem ser utilizados para os fins para que foram originalmente coletados, o que destrói de um só golpe a maior parte do seu valor. Quase todas das mais inovadoras e importantes utilizações da aprendizagem automática envolvem precisamente utilizar os dados para fins que não foram antecipados por ninguém, como por exemplo utilizar as pesquisas na Web para detetar o princípio duma epidemia. (Para já não falar de todas as grandes descobertas científicas que foram feitas por acidente, como a penicilina e os raios-X.) Não satisfeito com isto, o RGPD exige ainda que os algoritmos forneçam explicações das suas decisões, aparentemente desconhecendo que os algoritmos de aprendizagem de maior precisão, como o deep learning, são tipicamente os que não são capazes de o fazer. Eu prefiro ser diagnosticado por um sistema que tem precisão de 99% e não fornece explicações do que por um que tem precisão de 90% e as fornece. Outras pessoas poderão ter a preferência oposta, mas a decisão deve ser tomada por cada um de nós, não imposta arbitrariamente pela lei. E como se isto não chegasse, o RGPD inventa ainda um novo direito: o direito a ser esquecido, ou seja, o direito a exigir que dados sobre nós sejam apagados. Soa bem, mas o meu direito a ser esquecido infringe o teu direito a recordar-me, e que espécie de inteligência podem os computadores ter se a sua memória pode ser arbitrariamente mutilada? Chegados a este ponto, esperaríamos que a UE se desse conta do seu erro e revogasse o RGPD, mas não: pelo contrário, passou de seguida a Lei dos Mercados Digitais e a Lei dos Serviços Digitais, para agravar ainda mais o disparate, e vem já a caminho a Lei da Inteligência Artificial, que seguramente ganhará lugar de honra como uma das mais mal concebidas da história da UE.

A IA E O DESEMPREGO

E a lista de perigos imaginários da IA continua, mas concluo com um dos mais salientes: a onda de desemprego que supostamente vai provocar. Segundo esta teoria, a Revolução Industrial automatizou o trabalho manual, a IA vai rapidamente automatizar o trabalho intelectual, e quando isso acontecer nada sobrará para os trabalhadores humanos. Mas a Revolução Industrial não levou ao desemprego em grande escala: pelo contrário, criou muito mais empregos do que eliminou, e o mesmo acontecerá com a IA. Cada emprego consiste em múltiplas tarefas, e tipicamente a IA é capaz de desempenhar algumas mas não outras. O importante é que cada trabalhador perceba para que partes do seu trabalho pode utilizar a IA, e ao fazê-lo aumentará a sua produtividade e acabará por ganhar mais, não menos. A concorrência não é entre homens e máquinas, mas entre homens com máquinas e homens sem elas, e os primeiros vão inevitavelmente ganhar. Economicamente, o efeito principal da IA é diminuir o custo da inteligência. E quando o custo diminui, a procura aumenta. Com a IA, coisas como a medicina, o direito, o jornalismo, a programação e a educação poderão ser feitas a uma escala vastamente maior que hoje, com todos os benefícios e empregos adicionais que isso implica. Em particular, quando o preço dum produto baixa, o valor dos produtos complementares sobe; e precisamente o grande e indispensável complemento à IA é a inteligência humana. Mas, dizem os alarmistas, e os trabalhadores que não são capazes de se retreinar para novas ocupações? Seguramente esses vão ser um grande problema? Não. Quando os preços baixam devido à automatização — e a IA não é mais que uma forma de automatização — sobra mais dinheiro aos consumidores, que por sua vez o gastam em, por exemplo, jantar fora mais ou comprar uma casa melhor, dando emprego a cozinheiros, trabalhadores da construção civil, e muitos outros. O mundo, e os países desenvolvidos em particular, precisa urgentemente de aumentar a produtividade para compensar o envelhecimento da população. A IA é a melhor solução que temos para este problema. Queremos retardá-la ou acelerá-la?

OS RISCOS REAIS E OS BENEFÍCIOS

Um dos problemas com a atual obsessão com riscos imaginários da IA é que desvia a atenção dos riscos reais. Um dos maiores é o da sua utilização por regimes autoritários como a China para a repressão e para fins militares. Mas nós não podemos impedi-los de o fazer; a única solução é desenvolvermos a nossa IA melhor e mais rapidamente que eles, utilizando-a para melhorar o funcionamento da democracia, defender os direitos humanos, e derrotar esses regimes nos campos de batalha reais e virtuais. Do mesmo modo, a melhor forma de combater o uso da IA por criminosos é assegurar que a polícia tem melhor IA. Os assaltantes a bancos fugirem de carro não é razão para proibir que os carros sejam mais rápidos que os cavalos; é razão para a polícia ter carros mais rápidos que os assaltantes. Mas o maior risco da IA é a estupidez artificial. Não conheço um único caso até à data de danos causados por IA demasiado inteligentes. Os danos causados por IA demasiado estúpidas, por outro lado, são incontáveis. Todos os dias, as IA tomam milhões de decisões, desde a concessão de crédito à escolha de candidatos para empregos, mas como não têm senso comum, cometem erros desnecessários. A tarefa mais urgente para nós é dotar as IA desse senso comum, e o alarmismo tolo não ajuda nada.

Acima de tudo, os potenciais benefícios da IA são extraordinários. Curar as doenças, gerir melhor as cidades e o ambiente, gerar riqueza, fazer novas descobertas científicas, dar a cada um de nós uma infinidade de assistentes pessoais que nem as pessoas mais ricas do mundo podem hoje ter: quanto mais cedo, melhor. O maior risco da IA é que a obsessão com os seus riscos nos leve a perder esta oportunidade sem par.