“Quem tem medo da IA”

Eugénio Oliveira

Comunicação oral na sessão “Separação entre a Consciência Humana e a Inteligência Artificial” promovida pelo GAF- Grupo de Ação Filosófica da Universidade do Porto, 18 de maio de 2017.

0- O que é a IA ?

A Inteligência Artificial (IA) é todo o processo bqseando um estudo (se falamos de uma área científica em si) ou o produto desse estudo (se nos referirmos ao seu objectivo). E esse objectivo é, no caso da IA, o desenvolvimento de entidades computacionais capazes de exibir comportamentos semelhantes aos humanos em atividades que requerem aplicação de inteligência.

“A **Inteligência** pode ter muitas **facetas** como a **criatividade**, a resolução de **problemas** complexos, o reconhecimento de **padrões** [em dados, texto ou imagem], **classificação** de situações, **aprendizagem** de regras de raciocínio, indução, dedução, construção de analogias, optimização, sobrevivência em um ambiente (**adaptação**), **compreensão** e processamento de linguagens, extracção de conhecimento e muito mais….

No verão de 1956 na reunião do **Dartmouth College** no New Hampshire, John McCarthy propôs que fosse dado financiamento para **10 investigadores** fazerem um estudo de **2 meses** para que se debruçassem sobre a nova área de IA.

O estudo baseava-se na “**conjectura** de que toda e qualquer **característica** da inteligência poderia ser, em princípio, **descrita** com tal precisão que uma máquina poderá ser construída para a **simular**.

Nas décadas que se seguiram forma muitas as direcções de investigação que se lançaram em IA. A principal foi a de algoritmos de **Resolução de Problemas** (jogo de Xadrez, optimização de percursos para um robô, …)

baseados na abordagem **SIMBÓLICA** incluindo as lógicas. Os métodos de **Representação do Conhecimento e a Aquisição** desse Conhecimento foram as principais realizações.

Relativamente a métodos de **Aprendizagem** (para estreitarmos o foco) distinguem-se **5 tribos**: Os simbolistas (usando lógica e algoritmos formalizados) os conecionistas (mimando parte do cérebro ao nível do neurónio), os evolucionistas (tentam melhorar soluções por processos semelhantes à evolução natural (Seleção, cruzamento, reprodução) os bayesianos (usando redes de dependência probabilística) e os "analogizers" que incluem outro tipo de algoritmos para classificação como os usados nos sistemas de Q&A.

Curiosamente tem sido muito **mais fácil** desenvolver sistemas computacionais para resolver problemas que são **difíceis para o homem** que aqueles em que ele tem melhor desempenho (p.ex. **reconhecimento** de milhares de faces diferentes ou compreensão da **linguagem** natural). O **senso comum** é extremamente difícil de reproduzir.

Até há um **lustre** atrás eu diria que , como Andrew Ng conhecido professor em Stanford, é uma **distracção** desnecessária estarmo-nos a preocupar com uma possível **super-inteligência artificial** pois isso seria equivalente com a preocupação sobre a futura sobrepovoamento de **Marte**.

No entanto nos últimos tempos aconteceram dois fenómenos de **rutura** na evolução paulatina dos sistemas ditos de IA que nos devem levar a tomar desde já **precauções**:

Um foi o desenvolvimento do método a que se chama **Deep Learning** que, basicamente tem obtido grande sucesso na interpretação por etapas cada vez mais **abstractas** de grandes quantidades de dados (**petabytes** 10^{15} bytes). Reconhecimento de **imagens**, interpretação de **LN**, são realizações que se tornaram efectivas devido ao DL)

O outro foi precisamente a enormíssima quantidade de **dados** que tudo e todos estamos em todos os momentos a produzir sempre que clicamos, nos

ligamos e até os objectos e locais produzem, nos supermercados, nas ruas, etc (o advento do **IoT**).

Esta **combinação é explosiva** e, por falar em explosões, ainda há uns meses num painel em que colaborei na Califórnia, se dizia que os cientistas nucleares sabiam que tinham subestimado a importância futura da energia nuclear e quando reagiram era tarde e o mundo passou a ser governado muito à base desse resultado científico.

Hoje para além dos **grandes êxitos** badalados da IA ela é quase **invisível** como permitir ao **Facebook** ou ao **tweeter** decidir quais os updates ou os tweets a mostrar a cada utilizador. Permitir que grandes companhias conheçam o **perfil** dos seus clientes como nunca antes acontecera, a gestão inteligente de **satélites** ou as tentativas de descodificação de sequencias de **ADN** (PDomingos) ou da detecção de **carcinomas**. Muitos agentes **BDI** jogadores na **bolsa** conseguem grandes resultado.

Padrões de fala e de escrita analisados pelos novos sistemas cognitivos, como o IBM **Watson**, irão fornecer sinais fidedignos de um estágio inicial de **doença mental** e neurológica, o que pode ajudar médicos e pacientes a melhor prever, monitorizar e acompanhar estes distúrbios.

Na IBM, os cientistas estão já a usar áudio e **transcrições** de **entrevistas psiquiátricas**, juntamente com técnicas de machine learning processamento de linguagem natural, para encontrar padrões de fala e ajudar os médicos a prever e monitorizar alguns tipos de psicose, esquizofrenia, mania e depressão. Hoje em dia, bastam **300 palavras** para detetar a probabilidade de um paciente vir a sofrer de algum género de **psicose**.

Por isso por todo mundo, inclusive a **Casa Branca** antes, discute os possíveis **impactos** de uma IA geral (não só específica e estereotipada como Diagnóstico Médico) ou "**Strong** AI".

1) Pode o teste de Turing provar strong AI?

É verdade que um robô numa **fábrica** de automóveis pode ser muito útil sem ser inteligente. É verdade que o **Deep Blue e o AlfaGo** podem ganhar a um campeão mundial respectivamente de Xadrez ou de GO **sem ter consciência** disso nem senso comum.

Toda a gente já ouviu falar do **Teste de Turing** que classificaria como inteligente o computador que, estando numa sala com humanos, conseguisse ludibriar outro humano noutra sala que sem ver e só por comunicação (por e. escrita) não o identificasse como computador.

A resposta a se este teste seria **efectivo** ou não **depende** de algumas precisões como por ex: Quanto tempo deve durar a interacção? Que **perguntas** são feitas? Qual o grau de **conhecimento** de quem pergunta?

John Searl avançou com o famoso argumento da "Sala Chinesa "**Chinese Room**". E conclui que mesmo uma máquina inteligente que tivesse um programa capaz de manipular, passo a passo, símbolos chineses e encontrar um significado para toda uma sequência deles, retribuindo como output outros caracteres chineses que parecessem respostas ao input, nós não poderíamos dizer que a máquina compreendesse a língua chinesa (mandarim) pois **processava os símbolos sem compreendê-los** e sem intencionalidade. E concluía que a **IA Forte seria Falsa**.

Mas isto **não é um argumento, é um paradoxo** tal como o paradoxo de Zenão.

Zenão tentava argumentar que o rápido Achiles nunca ultrapassaria a tartaruga. A estratégia argumentativa foi a de decompor a situação em partições tal que o evento (ultrapassagem) não aconteceria. Ou seja em pequenas distâncias que permitiriam que enquanto Achiles tentava chegar à tartaruga ela já se tinha movido mais um pouco. E como o evento não se verifica nas partições também não se verifica na situação total.

Ora isto é um paradoxo e não um argumento pois contradiz factos observáveis.

O mesmo acontece com o **pseudo-argumento de JS**.

Ele **particiona o diálogo** com o computador em pequenos **passos** correspondentes a **instruções** do computador onde apenas há manipulação de símbolos, nenhuma compreensão real e, portanto, também para a **operação completa também não existe compreensão** nem **consciência** de saber chinês. Mas se calhar a compreensão é **precisamente o conjunto desses pequenos passos** elementares.

Jean E. **Tardy** o autor de vários livros como o The Meca Sapiens Blueprint dizia "Mas então não existe movimento nos **filmes** ?" Se eu partir os filmes em **frames**, em nenhuma delas nós observamos movimento. São estáticas. Mas o conjunto, a combinação delas dá o movimento.

A sala chinesa é um paradoxo esperando por ser contradita no próximo futuro?

Não podemos ser tão definitivos. O conceito de **consciência" não está definido sem ambiguidades**. Consciência igual a autoconhecimento (de si próprio)? E o "próprio" como se define? O conjunto de experiência de cada indivíduo incluindo sensações e raciocínio?

Um **bom TT** poderia indicar uma proximidade **assimptótica** com um certo tipo de inteligência.

Mas eu até preferiria o **Teste de Durkheim**, o sociólogo. Seria provar que a nossa entidade computacional saberia **participar em conjunto e seguindo as leis sociais**, na resolução de problemas complexos. Muito mais conhecimento, incluindo do **senso comum, será necessário** para cooperar e competir num grupo.

Há várias **mentes específicas de um domínio** (como jogar Xadrez ou condução autónoma na estrada) não incluem consciência, livre arbítrio, etc.

O bom teste de Turing não se limitaria a detetar **inteligência como a humana mas também comportamento como** o humano em sociedade

2) Funciona a mente como um computador?

Inteligência e Autonomia são duas propriedades que são definidoras do ser humano. O facto da IA se propor inclui-las nos sistemas faz pensar numa **ameaça** aos humanos. Tem-se a perceção que sistemas de IA e Robótica poderão tomar conta de muitos dos nossos **empregos** e das nossas qualificações. (VD)

Porque os sistemas de **IA tomam decisões (pensam) e podem interagir (texto ou voz) há a tendência de os comparar às pessoas**. Mas as **capacidades** e habilidades ainda tem focos diferentes. Os **humanos** são rápidos em processamento **paralelo** como por exemplo no reconhecimento de padrões (caras, p.ex.) e mais **lentos** no processamento **sequencial** (como o raciocínio **lógico** onde temos uma cadeia de inferências) os **computadores** apenas conseguem funcionar em paralelo para domínios estreitos mas são **super-rápidos na computação sequencial**. Ou seja a forma de "raciocinar" é **diferente da humana** (mas **submarinos** não nadam nem **aviões** batem as asas e são melhores que homens ou pássaros).

O que é a mente? E a que tipo de **entidade computacional** nos estamos a referir?

O que é a mente? Uma **propriedade emergente do cérebro** que dá aos humanos um conjunto de faculdades cognitivas incluindo inteligência, consciência, livre arbítrio, raciocínio, memória emoções, etc.

Aconselho o livro recentemente publicado "**The Digital Mind"** do meu colega Arlindo Oliveira do IST.

Os computadores ainda são de **silício** e é muitíssimo diferente tentar mimar um **cérebro "in silico" e "in vivo".**

A competição **Jeopardy** ganha aos humanos pelo programa **Watson** que, em voz sintetizada, respondia a questões difíceis. Mas eram sobretudo Factos históricos (**factóides**).

JSearle diz que Watson não compreende as perguntas nem as respostas, apenas manipula símbolos. Nem sabe se ganhou a competição. Talvez sim porque se esforçou por ganhar o jogo.

Ficou feliz por ganhar? Talvez não. Mas eu acho que o poderia programar para entrar num estado **emocional** semelhante à felicidade. Tal seria reconhecido porque lhe iria alterar a maneira de raciocinar, agir e memorizar durante certo tempo, como se estivesse alegre ou feliz ou então ansioso ou com medo. Isso eu acho que sei fazer.

Fazer o **download de um cérebro** para uma entidade artificial (robô, computador, rede de computadores) necessitaria de uma **Engenharia Reversa** que está muito longe de ser feita em pormenor. Os atuais métodos baseados em **MRI s**ão muito superficiais. Portanto a chamada Whole Brain Emulation) é para já inalcançável.

Não conseguimos ainda detetar a informação estrutural em detalhe. A resolução das **atuais técnicas de análise do cérebro é de 1mm^3** e dentro dele existem entre 50 e 100 mil neurónios cada um com centenas ou milhares de sinapses (cada sinápse do tamanho de 20 a 200 nanómetros. Reproduzir in silico o que existe in vivo será difícil.

E já imaginaram as implicações de algumas falhas na cópia?

Também poderíamos tentar **fazer evoluir um cérebro digital.**

Se conseguirmos simular a evolução de um cérebro teremos sistemas inteligentes neuromórficos que poderão levar à emulação digital de um cérebro.

A evolução de um cérebro digital necessitará de imensos **estímulos** cada vez mais complexos o que exigirá muitos **sensores** e um corpo . Muito tempo

seria necessário para uma simulação em tempo real incluindo interacção com outros humanos. Mas pode ser um caminho.

A resposta para já é não!!!!

3) É possível strong AI alcançar auto-consciência? É a consciência um requisito necessário para strong AI?

4) Será que podemos alcançar strong AI num futuro próximo?

Se existirem mentes artificiais de **inteligência geral coloca-se o problema da consciência**. Ganhamos consciência de nós ao **acordar** e perdemo-la ao adormecer.

Não vou aqui discutir a opos**ição entre Dualistas e monistas:**

Dualistas dizem que temos **duas "realms" (domínios**) O **físico** que é bem compreendido e o **não físico onde a consciência de nós existe** e que interage (como?) com o primeiro. As **religiões** exceto o budismo são baseadas neste dualismo. O **dualismo Cartesiano** propõe que cérebro e mente são duas coisas diferentes. A glândula pineal controlaria e permitia a interacção. Claro que **não há qualquer evidência do dualismo** e desta interação.

Monismo de (Christian von Wolff)

Desde o sec. XIX que Hobs defendeu que toda a **experiencia humana reside nos processos biológicos contidos no corpo (incluindo ao nível genético).**

Aliás os Cientistas atuais são quase todos materialistas sabendo que há apenas uma realidade que gera todos os fenómenos incluído consciência e o conceito de si próprio.

Há quem considere que a **consciência** não é assim tão inatingível (como o Searle pensava) mas sim que ela **poderá emergir da actividade coordenada de muitas funcionalidade e mecanismos mais simples.**

Portanto será possível no futuro a consciência de entidades artificias. Mas a **IA forte pode existir sem isso** porque podemos fazê-los **cooperar na resolução de problemas muito complexos, aplicar leis socio-éticas (na condução autónoma por exemplo) e dotá-los de estados emocionais elementares**. Por ex. se algo de assustador acontece tal pode ser reconhecido e o estado interno do sistema alterado (como se tivesse medo) despoletando comportamentos adequados (abandonar planos de acção e substituindo-os por outros).

IA Forte até um certo ponto sim, proximamente. Pensando sobretudo no **comportamento** exterior. Mas se pensarmos em **mimar completamente cérebros e mentes, não vejo** essa possibilidade no futuro próximo.

5) Quais seriam as implicações sócioeconómicas do alcance de strong AI?

A **motivação** para o conseguir é enorme. Para falar em **economês**, actualmente no mundo o PIB pode duplicar em 5 anos e, com estas novas tecnologias reproduzindo entidades inteligentes poderia dobrar em semanas. Não subestimemos a **ganância** das sociedades …

A pergunta sobre a implicação na **Ética** seria mais importante.

Eu tenho preconizado que os sistemas decisores deverão ter sempre "**The H in the Loop"**. Especificados de tal forma que os sistemas tenham **ART em ARTificial Intelligence**: Accountability, Responsibilty, Transparency

"**Accountability**": Isto é a **quem nos devemos dirigir se um automóvel auto-conduzido atropelar um peão**? Ao **construtor** do hardware do veiculo, dos **sensores** e atuadores? Ao desenvolvedor do **Software** que implementa o sistema de tomada de decisão? Às **autoridades** que permitem que tais veículos circulem nas estradas? Ao **condutor** que personaliza a tomada de

decisão automática? Ao próprio **carro-robô** pois o seu comportamento também é ditado pelo que foi aprendendo com a experiência? A **todos** Eles? (VD)

"**Responsibility**": Os sistemas de IA deveriam ter a responsabilidade de tornarem **claras e compreensíveis as suas decisões.** Não é o que se passa actualmente com os sistemas de "**Deep learning**" que herdam muito do paradigma das redes Neuronais Artificias e que, portanto, ao contrário dos SBC (KBS) são como caixas pretas onde entram dados e saem conclusões.

"**Transparency**" tem a ver com a especificação, desenvolvimento e **reprodutibilidade** dos sistemas de IA. Tal implica a compreensão do funcionamento e, eventualmente, a decisão quanto à automatização completa ou à preferível inclusão do "Human in the Loop" que eu advogo na maioria dos sistemas que já desenvolvi.

Não cremos que alguma vez os **humanos ficarão obsoletos** mesmo que haja transferências de competências (tal como aconteceu na **Revolução** industrial).Por ex. a **UBER** teve de empregar muitos especialistas nos veículos de autocondução (p.ex. só 50 vieram do Instituto de Robótica da CMU). **Especialistas** em IA são muito procurados em **Wall Street.** Tarefas mais mecanizadas (mesmo baseadas em conhecimento) serão mais rapidamente automatizadas mas seria útil **manter "the Human in the** loop" para assegurar bom senso, preocupações sociais e por vezes intuição nas máquinas.

Mas estamos num **Hype da IA e isso é perigoso**.

Nos últimos anos o **Graal** é "*Agora não temos de programar os computadores. Eles programam-se a si próprios*"
O outro Graal é o chamado **Algoritmo Mestre** (como a chave mestra que abre todas as portas).
(P Domingos). A tese central do seu livro The MA:
All knowledge—past, present, and future—can be derived from data by a single, universal learning algorithm."

O Chatbot do Twitter criado pela Microsoft **TAY** tornou-se incrivelmente racista, xenófobo e nazi pois foi alimentado com frases que traduziam essas ideias ("Hitler did nothing wrong").

Fotos com crianças nuas a fugir dos bombardeamentos no Vietnam (e que ganharam prémios internacionais) são automaticamente **banidas** do Youtube e Facebook por programas que as consideram imorais. E é o contrário!

Tomar decisões não connosco mas por nós é errado.

A **substituição de empregos** existirá mas as sociedades como um todo **recompõe**-se e ultrapassam as revoluções económicas para novos patamares. Mas há sempre muitas **pessoas** que podem vir a ser **trituradas** no processo e é absoluto **dever** de todos não permitir que tal aconteça, chame-se isso reeducação, solidariedade ou, menos interessante, caridade.

- Sempre que possível manter "**The Human in the Loop**" evitando a total automatização.

- **Privacidade** de dados pessoais e **anonimização** de dados agregados tornados público

- Desenvolvimento da **inteligência a par com outras componentes** do comportamento humano como os estados emocionais.

O que torna a inteligência mais evidente é o reconhecimento de que as decisões tomadas tem em conta um certo **bom senso comum** a nível individual e social (o que não está bem definido. Talvez ainda se acreditasse que a terra fosse plana se nos baseássemos no senso comum). Também pode ser que o reconhecimento do papel que a emoção tem na própria razão ajude a fazer melhor IA mais de acordo com os valores humanos

A lei normalmente move-se mais lentamente que a tecnologia. Vai demorar bastante antes que alterações realmente significativas na lei permitam, por exemplo, o uso alargado da condução automática.

Mady **Delvaux** na sua tentativa junto da UE propôs em janeiro passado uma peça de **legislação detalhada que incluía dar uma cartilha de deveres e direitos civis à IA.** Tal incluía dar a robôs inteligentes uma **"e-personalidade"** limitada comparável ao que se faz com corporações. Um estatuto legal que permite a empresas processar e ser processada (pelo menos no respeitante a compensações).

Seja como for, não é cedo para clarificar posições sobre o potencial impacto dos sistemas baseados em IA na sociedade.