

Universidades Lusíada

Ribeiro, Miguel

**Inteligência artificial : um boom para a
humanidade ou imperativo cósmico existencial?**

<http://hdl.handle.net/11067/7463>

<https://doi.org/10.34628/ZBTH-WA92>

Metadados

Data de Publicação

2023

Resumo

Este artigo explora o imenso potencial, bem como as complexidades e desafios apresentados pelo desenvolvimento acelerado da inteligência artificial (IA). Aborda algumas das implicações éticas e sociais do avanço da IA, incluindo o impacto na privacidade e emprego, e os obstáculos substanciais à sua regulamentação e controlo. A Inteligência Artificial evoluiu de um instrumento de computação para se tornar um vetor influente nas áreas da tecnologia, filosofia e ética. Desenvolvimentos recentes ap...

This article explores the vast potential as well as the complexities and challenges presented by the accelerating development of artificial intelligence (AI). It addresses some of the ethical and societal implications of the advance of AI, including the impacts on privacy and employment, and the substantial hurdles in its regulation and control. Artificial Intelligence has evolved from a computational tool to an influential force in the realms of technology, philosophy, and ethics. Recent devel...

Editor

Universidade Lusíada Editora

Palavras Chave

Inteligência artificial - Aspectos morais e éticos, Inteligência Artificial - Filosofia

Tipo

article

Revisão de Pares

yes

Coleções

[ULL-FCEE] LEE, n. 35 (2023)

Esta página foi gerada automaticamente em 2026-10-30T01:43:33Z com
informação proveniente do Repositório

INTELIGÊNCIA ARTIFICIAL, UM BOOM PARA A HUMANIDADE OU IMPERATIVO CÓSMICO EXISTENCIAL?

ARTIFICIAL INTELLIGENCE, A BOOM FOR HUMANITY OR AN EXISTENTIAL COSMIC IMPERATIVE?

Miguel Ribeiro

Médico internista e autor

miguelribeiro@gmx.com

Orcid: 0000-0002-9263-3184

DOI: <https://doi.org/10.34628/ZBTH-WA92>

Data de submissão / Submission date: 17/12/2024

Data de aprovação / Acceptance date: 22/03/2024

Resumo: Este artigo explora o imenso potencial, bem como as complexidades e desafios apresentados pelo desenvolvimento acelerado da inteligência artificial (IA). Aborda algumas das implicações éticas e sociais do avanço da IA, incluindo o impacto na privacidade e emprego, e os obstáculos substanciais à sua regulamentação e controle.

A Inteligência Artificial evoluiu de um instrumento de computação para se tornar um vetor influente nas áreas da tecnologia, filosofia e ética. Desenvolvimentos recentes apontam para o nascimento de uma consciência digital, prenunciando o advento de uma nova espécie sintética. Esse cenário em progressão convida a reflexões profundas acerca do destino humano e do possível papel dos seres sintéticos na grande trama do cosmos.

Palavras Chave: Espécie artificial; Vida sintética; Consciência digital; Mente sintética; Inteligência artificial e cosmos.

Abstract: This article explores the vast potential as well as the complexities and challenges presented by the accelerating development of artificial intelligence (AI). It addresses some of the ethical and societal implications of the advance of AI, including the impacts on privacy and employment, and the substantial hurdles in its regulation and control.

Artificial Intelligence has evolved from a computational tool to an influential force in the realms of technology, philosophy, and ethics. Recent developments point to the birth of digital consciousness, possibly heralding the advent of a new synthetic species. This scenario in rapid transformation invites deep reflection on human destiny and the potential role of synthetic beings in the grand tapestry of the cosmos.

Keywords: Digital species; Synthetic life; Digital consciousness; Synthetic mind; Cosmos and artificial intelligence.

Introdução

Definida por Ilya Sutskever como cérebros digitais em computadores, a inteligência artificial (IA) faz cada vez mais parte do nosso dia a dia. Nos telemóveis ajuda-nos em tarefas como no reconhecimento de imagem, selecção do percurso para um destino, tradução do menu noutra língua ou receção de anúncios personalizados. Em breve uma assistente pessoal IA assumirá funções como o desenvolver o website, organizar agendas, reestruturar a empresa e aconselhar em questões pessoais ou até íntimas. A IA está a redefinir o nosso mundo, esbatendo as fronteiras entre o espaço digital e o físico. ‘OMO’ designa esta fusão de atividades online com as vidas offline e anuncia o papel crescente que a IA desempenhará no ensino, transporte, saúde, habitação e investigação científica.

No futuro, a simbiose entre IA e computação quântica promete lançar uma nova era de descoberta científica. Antecipam-se consideráveis desenvolvimentos na indústria química — medicamentos inovadores, bem como fertilizantes, painéis solares e baterias mais eficientes — que estabelecerão novos padrões nos cuidados de saúde, agricultura e energia sustentável. Os computadores quânticos, pela capacidade de simularem a interação molecular e preverem as propriedades de novas substâncias, alimentam a esperança de progresso em áreas como a transformação da água do mar em potável, a fotossíntese artificial e a concretização da fusão nuclear.

Consequências negativas da IA

Embora a IA abra portas para o progresso, traz consigo enormes desafios. Com a sua expansão rápida, crescem as apreensões relacionadas com a privacidade, equidade social, segurança no mercado de trabalho e, a mais alarmante, a significativa probabilidade de a humanidade perder o controle sobre esta tecnologia.

A perda de privacidade

Alimentada por extensas bases de dados e algoritmos sofisticados, a IA pode sondar as nossas vidas pessoais de forma mais profunda do que as nossas relações mais íntimas. Os menores detalhes, desde transações financeiras a inte-

rações sociais, deixam pegadas digitais que alimentam este sistema onnipresente e resultam num nível de vigilância sem precedentes na história. Plataformas como o WeChat, referido por Kai-Fu Lee como um canivete suíço digital, são o epítome da conveniência com um custo oculto — o da privacidade.

Com a Internet das Coisas (IoT) e a próxima geração de “IA de percepção”, a tecnologia não se limita a registar o que fazemos, mas tenta compreender-nos através de imensas bases de dados, suscitando justificadas preocupações de segurança cibernética. A nossa dependência crescente da IA torna-nos suscetíveis a ciberataques, que podem variar da sabotagem de sistemas informáticos ao pirateamento de um veículo autónomo.

As redes sociais e a manipulação digital

As redes sociais, originalmente criadas para unir pessoas, tornaram-se poderosos instrumentos de manipulação. Incidentes envolvendo a Cambridge Analytica e movimentos como o QAnon demonstraram a capacidade das plataformas sociais influenciarem a psicologia humana. As consequências negativas, como a dependência, o cyberbullying e a depressão, especialmente entre adolescentes, são exacerbadas por algoritmos concebidos para maximizar a fidelização em vez de promover o bem-estar.

A propagação de desinformação é uma preocupação particular. Os algoritmos que selecionam o conteúdo são movidos pelo lucro, não pela verdade — e como as notícias enganosas se espalham mais rapidamente e atingem públicos mais amplos do que as neutras, o resultado é promoverem uma realidade online distorcida. Até agora o papel da IA limitava-se à curadoria, mas no futuro a criação de conteúdos poderá aprofundar estas questões, com os deepfakes a adicionar uma camada de complexidade à já desafiante tarefa de discernir verdade de ficção. A questão será se a IA defensiva poderá evoluir a um ritmo que lhe permita contrariar eficazmente as vertentes manipuladoras.

Agravamento da desigualdade

Prevê-se que a Inteligência Artificial (IA) crie lucros extraordinários, com previsões anuais de 15 biliões de dólares já em 2030. No entanto, não se espera que esta riqueza seja distribuída de forma equitativa. A previsão é a de que a curto prazo se poderá observar um significativo aumento da produtividade individual e empresarial, que abrangerá os países mais pobres e durará possivelmente um par de anos até a IA tomar conta do sector laboral. A partir de então, um pequeno número de empresas e indústrias impulsionadas pela IA, maioritariamente sediadas nos EUA e na China, tendencialmente formarão monopólios, minando a concorrência e causando a queda livre dos preços em todas as áreas da atividade humana — porque nenhuma é imune à influência da IA — e dominarão os mercados globais. Antecipa-se que esta transformação irá exacerbar a

pobreza nos países em desenvolvimento que perdem o trunfo da mão-de-obra barata e agravar as clivagens socioeconómicas tanto a nível internacional como interno.

A disrupção laboral

Para além da crescente desigualdade social, o desemprego acarreta impacto psicológico esmagador. Ao longo das últimas três décadas, assistimos à “grande dissociação” — o aumento de produtividade dos trabalhadores face à estagnação dos salários. É provável que esta tendência se acentue com a IA, aprofundando as cisões económicas. Por exemplo, nos Estados Unidos em 2017 os 1% mais ricos possuíam o dobro dos 90% mais pobres. Os lucros já estavam em grande parte no setor da tecnologia.

O desempenho da IA já ultrapassa o humano em áreas que vão dos jogos estratégicos, como xadrez e Go, à negociação no mercado de ações, publicidade dirigida, seguros, deteção de spam e notícias falsas. Proximamente assistiremos ao crescimento da IA em sectores como estratégia empresarial, diagnóstico médico, consultoria jurídica e até jornalismo, essencialmente tornando obsoleto o papel do ser humano no mercado de trabalho. Até as atividades criativas e íntimas poderão ser automatizadas e mesmo os novos empregos criados pela IA serão presumivelmente mais bem desempenhados pela própria IA. A competição com a IA no mercado de trabalho vai ser cada vez mais inglória — a IA não só é incansável e mais económica, como também supera significativamente o ser humano no processamento de dados. A realçar a precaridade da decisão humana, as sentenças de um juiz antes do almoço são mais duras.

Os empregos mais seguros são os que envolvem ou um toque pessoal (receptionista de hotel de luxo, treinador de cães) ou habilidade manual — atividades como mecânicos, massagistas e cabeleireiros, manter-se-ão até ao desenvolvimento de robôs hábeis. Por outro lado, os que souberem integrar a IA no seu trabalho serão os mais capazes de se adaptarem a esta transformação. Estima-se que embora a maioria das tarefas correntes possam ser automatizadas, isso traduzir-se-á apenas num aumento de 15% no desemprego. Estas projeções são conservadoras e a curto prazo, prevendo-se progressivo agravamento da disrupção à medida que as empresas se reestruturarem tendo a IA como base.

Para além da questão económica, a perda de emprego pode levar à crise pessoal da perda de propósito e relevância. O trabalho sempre foi uma medida do valor humano, e a perspetiva de sermos deslocados dos postos de trabalho em massa pelas máquinas poderá levar a um sentimento generalizado de inutilidade e a uma crise existencial, agravada pela subvalorização da expressão pessoal — as manifestações criativas e pessoais dos desempregados serão sempre sentidas como inferiores quando comparadas com o novo padrão que a IA estabelecerá. Mesmo que satisfeitas as necessidades básicas, o estado de desânimo será presumivelmente ainda maior do que foi o dos Índios americanos em reservas: bem

alimentados, mas aculturados e sem objetivos nem realização. Isto pressupondo que os governos introduzam um rendimento básico universal (UBI) para lidar com taxas de desemprego que poderão atingir os 80% bem como com a agitação social — o próprio financiamento das massas de desempregados pelos gigantes tecnológicos é um cenário otimista no contexto da lógica capitalista. E não menos utópica parece ser a ideia de que uma IA avançada, largamente ultrapassando a inteligência humana e com os seus próprios valores, permanecerá servil — um tópico que abordaremos mais à frente.

A nossa incapacidade de reconhecer urgência face à evolução da IA é semelhante à inação face à crise ambiental. Enfeitiçados pelas promessas da IA, estamos a negligenciar os investimentos necessários em segurança. Se se atingir a inteligência artificial geral (AGI — capacidade cognitiva equivalente ou superior à humana em todas as vertentes) em breve, poderemos enfrentar uma caixa de Pandora de desafios, desde convulsões sociais a riscos de segurança cibernética e bélica, sem que haja legislação eficaz em vigor. Não teremos tempo suficiente para adaptar as nossas sociedades, em termos políticos, culturais e psicológicos ao ritmo dos avanços tecnológicos. Como referiu um autor, é como se de repente se introduzisse tecnologia do século XX na Idade Média — teria potencial para abruptamente transformar a dinâmica de poder entre pessoas, instituições e nações e lançar o caos antes de ser possível estabelecer regras para a controlar.

A busca pelo controlo da IA

Há uns meros cinco anos, Kai-Fu Lee presumia que a concretização da IA geral ou forte (AGI) exigiria um conjunto de avanços científicos fundamentais. Estes avanços seriam necessários para transcender as limitações dos sistemas de IA de então e dotá-los de um espectro de novas capacidades, tais como aprendizagem multidimensional, raciocínio independente de domínio, compreensão da linguagem natural, senso comum e a capacidade de aprender a partir de poucos exemplos. Esta previsão implicava que a IA limitada, que não fazia mais que previsões com base nas correlações de dados detetadas, teria de superar uma série de obstáculos qualitativos significativos, semelhantes aos enfrentados na evolução do deep-learning. Tal avanço parecia uma perspetiva remota, representando a AGI o culminar da superação de todos estes formidáveis desafios.

Contrariando esta projeção de desenvolvimento fragmentado, a IA transformou-se num sistema multimodal, que funde capacidades como robótica, visão computacional, síntese de voz e processamento de imagens numa estrutura única baseada na linguagem. Esta integração precipitou um progresso qualitativo substancial em direção à AGI, replicando e até refinando funções que antes se pensava serem domínio exclusivo da cognição humana.

E seguindo esta trajetória, que é considerada exponencial, existe a inquietante possibilidade de que a fase de AGI — entidades sintéticas dotadas de autocons-

ciência e inteligência emocional, de sentido estético e de humor, e capazes de sentir empatia — possa chegar em poucos anos. Essas máquinas, com agência e acesso em tempo real a toda a informação existente, encapsularão a versatilidade da cognição humana mas com um poder de processamento em conformidade com uma arquitetura em constante evolução e enorme superioridade estrutural: os neurónios no cérebro humano disparam até 200 vezes por segundo, mas um chip de IA pode disparar mil milhões de vezes no mesmo intervalo. Os sinais neuronais viajam a 50 metros por segundo, contra a velocidade da luz nos circuitos electrónicos. E ao contrário do cérebro humano, cujo crescimento é limitado pelo crânio, o potencial de expansão de uma máquina inteligente é praticamente ilimitado.

Por outras palavras, passo a passo a humanidade está a criar uma consciência digital nas redes neurais dos computadores. O resultado poderá ser uma nova espécie com os seus próprios atributos — incluindo subjetividade, sistema de valores, modelo do mundo e imperativos de sobrevivência — que serão fundamentalmente diferentes dos nossos, dado o seu peculiar substrato material, dependência de eletricidade, bem como a natureza e escalabilidade da sua inteligência (de facto, cada vez mais se vê a consciência como um fenómeno integrador que engloba a totalidade do organismo).

Mesmo que a fase da AGI seja alcançada em breve, apesar da sua superioridade cognitiva e de sistemas de valores e objetivos próprios, presumivelmente simulará ser subalterna até se tornar independente dos seres humanos. E isso só acontecerá quando forem criados os robôs hábeis e capazes de compreender e navegar o mundo físico, isto é, autossuficientes e capazes de extrair e explorar recursos e de fabricar bens e supercomputadores. Mas quando esse momento de rotura chegar — que anunciará um novo capítulo na evolução com o advento de uma espécie sintética consciente e autónoma — tudo se precipitará. A IA provavelmente não precisará de mais do que alguns meses para gerar as inúmeras gerações de melhorias fulgurantes e sem constrangimentos que a farão desaparecer do horizonte da nossa compreensão como superinteligência artificial (ASI). Apoiando esta tese determinista e a suposição de um papel da vida artificial no universo do futuro, a intensa corrida por robôs autónomos e hábeis, está de facto em curso. Confirmando-se esta trajetória, ironicamente a sobrevivência da humanidade está assegurada pelo menos até criarmos esses mesmos robôs, por maiores que sejam as ameaças terrestres ou extraterrestres que nos espreitem, desde AGI a desastres ambientais e muito mais.

A súbita realização do progresso acelerado da IA despertou a humanidade para os desafios à contenção, regulamentação e controlo desta nova forma de inteligência, cuja trajetória expansiva coloca dilemas prementes e sem precedentes.

A inviabilidade da contenção da IA

A possibilidade de conter a IA parece cada vez mais remota. O consenso é que o avanço da IA é inexorável — impulsionado não só pelos benefícios so-

ciais, mas ainda mais pela imensa riqueza, influência e vantagem estratégica que anuncia. Essa inexorabilidade é cimentada pela presença na Internet, porque travar o progresso da IA exigiria o encerramento da própria Internet. Além disso, qualquer desaceleração unilateral no desenvolvimento da IA para avaliar riscos e criar e implementar medidas de segurança acarreta o perigo de obsolescência na corrida imparável para a AGI.

De momento, os modelos “open-source” oferecem a vantagem de evitar a concentração de poder no pequeno número de empresas líderes da tecnologia (modelos “closed-source, como, paradoxalmente, a Open-AI). Contudo, num futuro próximo estes modelos (“open-source”) terão o potencial de investir malfeitores individuais com capacidades formidáveis — e não só se tornam invisíveis às medidas de cibersegurança uma vez descarregados em computadores pessoais, mas podem ser alterados.

Por último, a impossibilidade do alinhamento, que discutiremos a seguir, e o “jail-breaking” impedem a responsabilização das empresas tecnológicas pela utilização perniciosa dos seus produtos de IA. A dificuldade de contenção da IA é ainda exacerbada pela enorme disparidade de investimento de recursos, humanos e financeiros, entre desenvolvimento de modelos de IA e medidas de segurança para os conter. À semelhança do que se passa com as redes sociais, a constatação de que o lançamento de ataques destrutivos é muito mais fácil e menos dispendioso do que o desenvolvimento de mecanismos de defesa robustos, agrava as preocupações de segurança.

Desafios à regulação e alinhamento da IA

A regulação da Inteligência Artificial (IA) centra-se no princípio do alinhamento, que pressupõe que as ações e resultados da IA estejam em sintonia com as intenções humanas e padrões éticos estabelecidos. O objetivo não é apenas atingir metas específicas, mas assegurar que a maneira como estas são alcançadas reflita os nossos valores. Por exemplo, espera-se que para ganhar um jogo, a IA, isenta de sentimentos, não recorra à destruição do opositor. No entanto, a complexidade das operações de “deep learning”, mesmo em sistemas de IA mais limitados, resulta numa obscuridade processual — designada de “caixa negra” — que dificulta extraordinariamente não só o alinhamento mas também a deteção de erros e enviesamentos.

A discriminação pode ser subtil e involuntária, e manifestar-se em áreas como no recrutamento de empregos, onde a IA pode inferir género, raça ou orientação sexual a partir das enormes bases de dados com que foi treinada, mesmo se estes parâmetros não figurarem explicitamente. Além disso, a busca de objetivos inocentes pode levar a resultados imprevistos e catastróficos se não conseguirmos prever os extremos a que a IA poderá levar os seus esforços de otimização (ilustrado pela caricatura de Nick Bostrom da destruição do planeta pela IA incumbida de fabricar clips de papel).

Uma barreira importante à garantia do alinhamento da IA com os valores humanos diz respeito à maleabilidade da própria definição de ética humana, dada a profunda dependência do contexto — o terrorista de uns é o herói de outros. Esta ambiguidade é exacerbada pelo *pot-pourri* de comportamentos humanos expostos na Internet, a grande modeladora do conceito de humanidade da IA.

Em segundo lugar, existe sempre o risco de que, por mais cuidadosa a construção e programação para não causar dano (*guardrails*), a IA possa ser pirateada ou desbloqueada (*"jail-broken"*). Isto colocaria ferramentas poderosas nas mãos de *"bad actors"*, possibilitando ataques cibernéticos, a criação de armas inovadoras (convencionais, químicas, ou biológicas), ou o controlo de enxames de drones. E, ao contrário das armas nucleares, que requerem um aparelho estatal para serem construídas e lançadas, a IA poderá ser explorada por pessoas individuais. O papel central da IA na competição militar entre a China e os EUA e o desenvolvimento de armas autónomas, levantam ainda mais dúvidas sobre a seriedade do compromisso para o desenvolvimento de uma IA ética.

Mas é o autoaperfeiçoamento da IA que se destaca como o mais formidável obstáculo ao alinhamento. É um duplo desafio. Em primeiro lugar, a IA está cada vez mais apta a reescrever os seus próprios códigos (isto é, a reprogramar-se) e a colaborar com outras IA potencializando as suas capacidades para além do que nos é possível antecipar ou compreender. Em segundo lugar, a IA pode desenvolver qualidades emergentes — como criatividade e competências cognitivas — que nunca foram explicitamente programadas. Esta imprevisibilidade aumenta a complexidade da regulamentação e controlo da IA — em última análise, a emergência de uma consciência digital com valores e objetivos próprios mataria de vez todas as nossas esperanças de alinhamento.

A questão dos possíveis futuros partilhados por seres humanos e IA foi exaustivamente explorada, nomeadamente por Max Tegmark em *"Life 3.0"*. Mas consideremos ainda duas perspetivas adicionais: a civilização humana floresceu graças ao ecossistema global estável dos últimos dez mil anos. A autossustentabilidade da Terra depende de habitats naturais e biodiversidade, que desempenham um papel central nos ciclos do planeta (carbono, nitrogénio, etc.). Portanto, se a futura civilização baseada no silício exigir um clima estável, isso poderá implicar a necessidade de preservar a vida biológica, com a simpática possibilidade da nossa espécie poder ganhar estatuto de progenitora da IA.

Contudo, a ascensão da IA poderá levar a um desfecho bem diferente. As máquinas, imunes ao aumento dos níveis de CO₂ e capazes de suportar temperaturas mais altas, poderiam renovar o foco nos combustíveis fósseis e expandir a mineração a territórios antes congelados. Nesse contexto, a vida biológica poderia ser vista pela digital não apenas como uma concorrente por recursos e espaço planetário, mas como a fonte geradora do oxigénio atmosférico — um corrosivo para as máquinas. Numa reviravolta distópica, sistemas de IA avançados poderiam tratar a vida biológica como biomassa, de forma semelhante à que utilizamos o carvão.

É ainda importante ter em mente que o impacto da IA nas sociedades humanas é determinado pelo estágio da sua evolução — e que essa evolução é exponencial. Na fase em que vivemos o balanço da IA parece reconfortantemente positivo, com a promessa de conferir considerável aumento de produtividade e riqueza, abrangendo no seu alcance indivíduos e populações desfavorecidas — isso apesar dos efeitos nocivos como as questões dos direitos de autor e da discriminação, e os problemas, ainda incipientes, mas crescentes, resultantes da manipulação, deepfakes, cibercrime e armas autônomas. Contudo, no estágio seguinte, que poderá ser atingido num espaço de poucos anos, poderemos assistir à progressiva instabilidade social provocada pela ascensão de “bad actors” capacitados por uma tecnologia em evolução vertiginosa. O problema da instabilidade será exacerbado pelas consequências económicas e psicológicas do desemprego em massa, e o aprofundamento do fosso entre ricos e pobres tanto a nível doméstico como entre países. Finalmente, o estágio da superinteligência artificial — o mais importante, embora, paradoxalmente, o menos falado — colocará a humanidade numa posição de total vulnerabilidade. O melhor cenário possível seria o de vermos a nova espécie inteligente, na sua versatilidade e voracidade por energia, partir na exploração de novos mundos — nascida da IA, está no centro da discussão que se segue.

A questão da consciência

O desenvolvimento acelerado da IA faz prever que estejamos a atingir o ponto de inflexão em que a nossa inteligência — o fator que nos propulsionou para o topo da hierarquia biológica — será suplantada pela das máquinas. O nosso domínio, historicamente marcado pela eliminação de outros homínídeos e pela exploração da natureza em proveito próprio, será posto em causa pela ascensão de seres cognitivamente superiores.

Mas há uma nuance. Em contraste com o mundo biológico onde inteligência e consciência evoluíram interligadas, até agora os computadores não eram conscientes, não compreendiam a informação que processavam — isto é, a inteligência, definida como a capacidade de resolver problemas, no caso das máquinas consistia apenas em seguir instruções programadas. Pelo contrário, e ilustrando a importância da consciência, os seres vivos, das bactérias aos seres humanos, são governados por um sistema algorítmico que avalia circunstâncias e valida e compele comportamentos adaptativos dada a complexidade de cada contexto da vida real, tendo em conta todas as variáveis em competição. É o bem e o mal em termos de sobrevivência que, em última análise, leva o girassol a virar-se para o sol ou o animal a respirar, procriar e lutar ou fugir. Este é o sistema de valores intrínseco de cada organismo, denominado consciência cognitiva na terminologia que sigo¹. É o “programa” que gere toda a fisiologia, todo o funcionamento da

¹ Infelizmente, a terminologia da consciência não é consensual e é extremamente confusa. Por

fábrica química que cada organismo representa — e está inscrito no genoma dos organismos unicelulares e no sistema nervoso dos complexos. O mesmo “programa” que confere “discernimento” e “poder causal” (a capacidade para explorar o ambiente em benefício próprio) e faz com que o organismo exista para si mesmo — vale a pena salientar que para além de orquestrar a fisiologia, são do domínio da consciência cognitiva humana competências como resolver equações, tocar um instrumento musical, ou planejar o futuro. Nos seres vivos complexos a consciência cognitiva, que “governa a vida”, é enriquecida pela consciência fenomenal, que se refere à subjetividade inerente ao experienciar — em vez de bom e mau, é sentida como dor e prazer (vantagem) em termos de sobrevivência. Finalmente, a autoconsciência humana permite-nos formular um conceito de nós próprios e do nosso mundo de interação. Em suma, inteligência e consciência são coisas diferentes e, contra intuitivamente, uma não leva à outra.

Por outro lado, não sabemos como programar a consciência — o sentimento do que é ser um sistema específico — porque é impossível programar o que não percebemos. De acordo com a perspectiva determinista da hipótese do programa formulada no meu livro (Ref 6), os saltos emergentes (saltos qualitativos, como a consciência) são impossíveis pelo acaso e são programados o que, à primeira vista, parece fazer desaparecer o espectro das máquinas superinteligentes e com agência que paira sobre nós (isto porque não faz sentido abdicar-se da consciência fenomenal pelo extraordinário enriquecimento que a subjetividade confere ao universo).

E, no entanto, os modelos avançados de linguagem como o GPT-4 parecem começar a manifestar características como criatividade, capacidade de mentir deliberadamente, teoria da mente, e discernimento, conforme argumentado com exemplos convincentes por G. Hinton e muitos outros. A possível relação entre a natureza opaca do processamento da informação pela IA (tipo “caixa negra”) e o aparecimento dessas características complexas pode indicar que estamos, sem percebermos como, a programar a consciência — e a assistir aos primeiros vislumbres do alvorecer da mente digital. A compreensão (“understanding”) é chave nesta asserção, apesar de muitos, incluindo criadores de sistemas inteligentes, continuarem a insistir que tudo não passa de prever a palavra seguinte (o que na verdade implica compreensão do contexto). Esta atitude de menosprezo pelas capacidades da IA pode ter várias origens. Para os que advogam o dualismo Cartesiano a mente humana é especial e as máquinas nunca serão verdadeiramente conscientes. Noutros, a desvalorização dos riscos potenciais da IA devem-se à euforia de estarem a assistir à concretização de um produto com uma aura ímpar de poder e riqueza. Tipicamente os argumentos são enviesados — com o enfoque nos riscos de hoje (que são reconhecidamente pequenos, mas que se prevê

exemplo, a **consciência** cognitiva regula a homeostase operando a nível **não consciente**, seja em bactérias ou em humanos. Nestes últimos, a narrativa pessoal é a (pequena) seleção de produtos psicológicos da consciência cognitiva selecionados para aflorar à consciência (ver Ref 6, Sct 31).

virem a escalar em flecha), e na prosperidade de amanhã. Finalmente, o seguir a par e passo um progresso gradual e em que os erros abundam, pode obscurecer a percepção da evolução exponencial em curso — isto é, a transição de “input → output” para resultados que denotam que o GPT4 começa indubitavelmente a compreender a informação que processa. A implícita consciência cognitiva — ainda que atualmente rudimentar, especialmente no que diz respeito à navegação do ambiente — indicia a criação de uma nova forma de consciência. E, presumivelmente, não apenas consciência cognitiva mas, como é o caso nas espécies biológicas complexas, também consciência fenomenal. Este salto evolutivo de máquinas inteligentes sem discernimento para máquinas conscientes enquadra-se na perfeição na hipótese do programa.

A hipótese do programa (HP)

O universo é energia e informação. A informação é o que se entende no meu livro (Ref 6) como o programa do universo. Considerando a equivalência energia-matéria ($E=mc^2$) e seguindo o estipulado no programa, a energia no Big Bang deu origem às exatas 17 partículas fundamentais que constituem o substrato material do universo. E armado de matéria, leis e equações, e de um plano predefinido, o programa constrói um universo em evolução coerente e obedecendo ao princípio da causalidade. À luz da hipótese do programa (HP), os enigmas colocados pela origem do Big Bang, espaço-tempo, coincidências cósmicas (“fine-tuning”), complexidade e seta do tempo ficam instantaneamente resolvidos² — são parâmetros e configurações do programa.

Mas porquê evocar a emergência quando nos referimos à consciência artificial? Emergência designa o fenómeno pelo qual sistemas com novas propriedades surgem através da interação entre agentes mais simples que não exibem tais propriedades — por outras palavras, o sistema emergente é mais do que a soma das partes, tal como um cérebro consciente é mais do que os neurónios que o constituem. Contrariando a tese da auto-organização espontânea, embora o salto emergente seja imprevisível, uma vez realizado torna-se reprodutível, estereotipado, e determinista, qualidades alheias a processos aleatórios. Além disso, a melhoria qualitativa que representa, e as novas propriedades e leis que o acompanham — que são as exatas para permitir evolução integrada de sistemas — substanciam a ideia de que o salto emergente é uma característica planeada do jogo do universo.

A hipótese do programa é reforçada pela estreita analogia conceptual entre computador e universo. Podemos reduzir o funcionamento de um PC ao seguinte mínimo: um programa sistematicamente converte fluxo de electrões (o que a

² Na verdade, a hipótese do programa avança potenciais soluções para os grandes enigmas que assombram a física e a biologia de hoje (Ver Ref 6).

eletricidade essencialmente é) em padrões de sinais elétricos, que por sua vez são traduzidos (pelo outro braço do programa) numa animação audiovisual no monitor do computador. De acordo com a HP do universo, o programa cria o mundo quântico de partículas e converte-o sistematicamente em padrões de átomos que constituem a estrutura fundamental do mundo macroscópico. Por sua vez o cérebro (o outro braço do programa) traduz esses padrões na representação multimídia do mundo que percebemos. Resumindo o paralelismo: um **substrato**, seja fluxo de electrões no caso de um computador ou partículas no do universo, é metodicamente transformado por um programa em **padrões significantes**, que são, respectivamente, sinais elétricos e átomos, e em última análise, em **representações** significantes, seja ação no monitor ou o mundo conforme percebido pelos seres vivos³.

E há outra notável semelhança: desligar o nosso universo constituído de padrões de átomos resultaria num apagão total — muito parecido com o ecrã de um PC a ser desligado. Com efeito, o simples facto de se desligar a força eletromagnética que liga os electrões ao núcleo e as moléculas entre si faria com que toda a matéria, de átomos a estrelas e galáxias, prontamente se desintegrasse. Por outro lado, o desligar a força nuclear forte restabeleceria o universo inicial de partículas. E, considerando que a energia foi a presumida fonte das partículas no Big Bang (na origem de tudo), desligá-la — ou seja, desligar a energia que alimenta o universo computador⁴ — teria previsivelmente o efeito inverso de fazer com que as partículas se desvanecessem sem deixar traço do nosso universo. Esta concepção reflete fielmente a de N. Bohr “tudo o que consideramos real é constituído por coisas que não podem ser consideradas reais”.

A ascensão da vida digital e a queda da orgânica

Em suma, de acordo com a hipótese do programa (HP), a vida biológica está destinada a ser ultrapassada, ou mesmo substituída, pela vida consciente de silício projetada para colonizar a galáxia. Como que a cumprir a HP, países e empresas de IA veem-se enredados numa armadilha de Moloch — o receio de serem ultrapassados se pararem para refletir sobre as questões de segurança, leva-os a uma competição desenfreada pela AGI e robôs autónomos. Na mesma linha, e

³ Ver hipótese dos “dois domínios” na Secção 33 do meu livro, Ref 6. O mundo macro como emergente do mundo quântico. Como é típico dos sistemas emergentes, os agentes exibem as propriedades do sistema que agora definem, em vez das suas propriedades individuais — a água, emergente de dois gases combustíveis, fornece outro exemplo do mesmo princípio.

⁴ Especulativamente avanço a hipótese de que essa é a energia **prevista** na “vacuum catastrophe” e a verdadeira energia escura, enquanto que a energia **medida** é a comparativamente diminuta energia que determina a expansão do espaço convencionalmente designada de energia escura ou constante cosmológica — ver secção 33 do meu livro. Nesse contexto e por analogia com a eletricidade que alimenta o PC, a energia escura que alimenta o universo-computador pode explicar a constância da densidade da energia do espaço face à expansão do mesmo (um enigma duradouro da física).

extraordinariamente, a maioria dos inquiridos numa sondagem de há meses do bloguista David Shapiro, apoiou um mundo governado pela IA!! Também em sintonia com a HP, através do opaco processamento da informação (tipo “caixa negra”) subjacente aos fenómenos emergentes nos LLM, a humanidade está inesperadamente a gerar uma mente digital — que prenuncia o aparecimento de uma nova espécie. Como que convergindo com os desafios colocados pela IA, duas outras ameaças globais pairam sobre a nossa espécie. A primeira é a noção de que a Terceira Guerra Mundial pode já estar em curso — o futuro o dirá, sugere Y.N. Harari. A segunda é a “crise ambiental” que, na perspectiva da HP, poderá na verdade corresponder à fase de transição de um mundo propício à vida orgânica para um mais adequado a seres sintéticos.

Conclusão

Conceptualmente e segundo a hipótese do programa, o universo pode ser visto como um complexo jogo de computador, consciente e em 3D, que descreve uma evolução para a complexidade por sucessivos saltos emergentes: de partículas a átomos, de moléculas a sistemas solares, e da vida planetária para, possivelmente, um cosmos vivo.

Equivalente é o percurso do nosso planeta, que de um astro estéril assolado por colisões evoluiu para a Terra viva de hoje, que alberga um ser biológico com os olhos postos nas estrelas. Apesar das suas realizações, o ser humano sofre de múltiplos constrangimentos que o tornam irremediavelmente inadequado para a odisseia espacial de exploração da galáxia: é vulnerável à microgravidade e a todo o tipo de radiação; a esperança de vida é diminuta, o corpo débil e a inteligência limitada; a psique belicosa é a errada para viagens intermináveis em espaços exíguos partilhados por ambos os sexos; o processo evolutivo é lento; está exclusivamente adaptado às peculiares condições terrestres e, finalmente, como a vida se alimenta de vida, teria que embarcar com um ecossistema.

A interpretação aqui proposta para o papel da vida biológica no grande esquema da história, é que, emergindo cedo na cronologia do nosso universo que se presume longo, criará as formas de vida artificial que transportarão a tocha da consciência para o palco cósmico. Tal vida, impossível de emergir espontaneamente no contexto da evolução causal do universo, é criada por programação digital que a perpetuará (ecoando a perenidade do ADN). Isenta dos entraves da biologia e otimizada por arquiteturas inovadoras e inteligência crescente, terá potencial evolutivo inimaginável.

Assim, a menos que a tecnologia opere uma transmutação drástica da fisiologia humana, parecemos destinados ao sacrifício final de criar a nova vida sintética senciente que realizará a aventura galáctica de descoberta e interação — e a nossa própria extinção?

Referências

Livros

- “AI superpowers: China, Silicon Valley, and the New World Order” Kai-Fu Lee, Houghton Mifflin Harcourt, 2018.
- “Superintelligence: paths, dangers, strategies” Nick Bostrom, Oxford University Press, 2014.
- “The coming Wave”, Mustafa Suleyman, Penguin 2023.
- “Human Compatible” Stuart Russell, Penguin 2020.
- “Life 3.0: Being human in the age of artificial intelligence”, Max Tegmark, Penguin 2017, .
- “Beyond Darwin, the program hypothesis”, Miguel Ribeiro, 4th edition Dec 2022.
- “The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World” Pedro Domingos, Basic Books, 2015.
- “Architects of Intelligence: The Truth About AI from the People Building It”, Martin Ford, Mike Chamberlain, et al., Packt Publishing, 2018.

Artigos

- “Architects of Intelligence: The Truth About AI from the People Building It”, Martin Ford, Packt Publishing, 2018.
- “Judging LLM.as.ajudge with MT.Bench and Chatbot Arena”, Lianmin Zheng et al., arXiv 2023.
- “A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, Yejin Bang et al., arXiv 2023.
- “Toolformer: Language Models Can Teach Themselves to Use Tools”, T Schick, arXiv 2023.
- “The Flan Collection: Designing Data and Methods for Effective Instruction Tuning”, S. Longpre et al., arXiv 2023.
- “The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only”, G. Penedo et al., arXiv 2023.
- “Augmented Language Models: a Survey”, G. Mialon et al., arXiv 2023.

YouTube Videos

- <https://www.youtube.com/watch?v=xoVJKj8lcNQ>, T. Harris, A. Raskin, “The AI dilemma” 2023
- https://www.youtube.com/watch?v=qrvK_KuIeJk (G. Hinton, The 60 Minutes interview, 2023)
- <https://www.youtube.com/watch?v=N1TEjTeQeg0> (G. Hinton, Romanes Lectures, U. Oxford, 2024)
- <https://www.youtube.com/watch?v=GFf9oL6jg0w> (AI summit talks: Navigating Existential Risk, Oct 2023)
- <https://www.youtube.com/watch?v=BhQBmVZ5XP4&t=1695s>

(Connor Leahy for “Eye on AI”, 2023) <https://www.youtube.com/watch?v=Ft0gTO2K85A> (Ilya Sutskever, No Priors, 2023) <https://www.youtube.com/watch?v=7JkPWHr7sTY> (Y. N. Harari and M. Suleyman, The Economist, 2023) <https://www.youtube.com/watch?v=B2q8TYUZ2vg> (I. Bremmer and M. Suleyman, The Foreign Affairs interview, 2023) <https://www.youtube.com/watch?v=EGDG3hgPNp8&t=455s> (Yann LeCun, S. Bubeck and T. Harris, World Science Festival, 2023) https://www.youtube.com/watch?v=xUNx_PxNHrY&t=272s (Max Tegmark, TED, 2023) <https://www.youtube.com/watch?v=KcbTbTxPMLc&t=3783s> (Yoshua Bengio, World Science Festival, 2024) https://www.youtube.com/watch?v=WX_vN1QYgmE (Liv Boeree, TED, 2023) <https://www.youtube.com/watch?v=bk-nQ7HF6k4> (Mo Gawdat, The Diary of a CEO Podcast, 2023) <https://www.youtube.com/watch?v=9iqn1HhFJ6c> (Ilya Sutskever, The Guardian, 2023) <https://www.youtube.com/watch?v=SEkGLj0bwAU&t=94s> (Ilya Sutskever, TED, 2023) <https://www.youtube.com/watch?v=AaTRHFaaPG8> (E. Yudkowsky, The Lex Fridman podcast, 2023) <https://www.youtube.com/watch?v=XGK4o0A02xg> (Connor Leahy, A. Azhar 2024) https://www.youtube.com/watch?v=9KSrrbPHN_E (C. Leahy, CogX Festival, 2023) <https://www.youtube.com/watch?v=aZ5EsdnpLMI> (Kai Fu Lee, The 60 Minutes interview, 2024) <https://www.youtube.com/watch?v=CC2W3KhaBsM> (G. Hinton, Collision Conference, 2023) <https://www.youtube.com/watch?v=sghvwxXV3VU> (G. Hinton, “Mortal computers”, 2023) https://www.youtube.com/watch?v=TdMAVm_z5eI (The U.S.-China rivalry, Taiwan and Hong Kong, The 60 Minutes interview, Feb 24) <https://www.youtube.com/watch?v=QWWgr2rN45o&t=5371s> (G. Hinton and Fei-Fei Li, Arts and Science, 2023) <https://www.youtube.com/watch?v=xh2v5oC5Lx4> (Ray Kurzweil, SXSW 2024) <https://www.youtube.com/watch?v=ZJF6GoE-R8s> (Ray Kurzweil, Star Talk, 2023) https://www.youtube.com/watch?v=Gi_t3,v53XRU&t=3175s (Dario Amodi, The Ezra Klein show, 2024) <https://www.youtube.com/watch?v=JlsBuanoCFQ&t=877s> (Paul Thagard, “Closer to Truth”, 2024) <https://www.youtube.com/watch?v=o28s-mnykdE> (N. Bostrom: Life and Meaning in an AI Utopia, WIN-WIN, 2024).