ORAÇÕES FLUP

ORAÇÕES DE SAPIÊNCIA

# PORQUE PRECISAMOS, AFINAL, DE UMA ÉTICA DA INTELIGÊNCIA ARTIFICIAL?

MARKUS GABRIEL

**FACULDADE DE LETRAS** DA UNIVERSIDADE DO PORTO
PORTO **2025**

U.PORTO
FACULDADE DE LETRAS
UNIVERSIDADE DO PORTO

Prof. Dr. Markus Gabriel doutorou-se em 2005 na Universidade de Heidelberg, onde obteve a habilitação em 2008. Pela sua tese sobre a filosofia tardia de Schelling, foi distinguido com o Prémio Ruprecht-Karls. Após ter sido Professor Assistente na New School for Social Research, assumiu em 2009 a cátedra de Epistemologia, Filosofia Moderna e Contemporânea na Universidade de Bona.

Desde 2012, dirige o Centro Internacional de Filosofia e, desde 2017, o Center for Science and Thought da mesma universidade. Desde 2020, é Professor Convidado e Co-diretor do Institute for Philosophy and the New Humanities na New School, em Nova Iorque. Em 2024 tornou-se Senior Global Advisor no Kyoto Institute of Philosophy e, em 2025, Professor Especialmente Nomeado no Kyoto University Institute for the Future of Human Society. Foi Professor Convidado e Fellow em várias universidades internacionais, incluindo a Universidade de Lisboa, a Universidade de Aarhus, a PUC-Rio de Janeiro, a PUCRS, a New York University, Berkeley, Stanford e Paris 1-Panthéon Sorbonne. Atualmente, é Gulbenkian Visiting Professor in the Humanities na Universidade do Porto.

As suas publicações mais recentes são *Sense, Nonsense, and Subjectivity* (Harvard University Press, 2024) e *Moralische Tatsachen* (C. H. Beck, 2025).

# PORQUE PRECISAMOS, AFINAL, DE UMA ÉTICA DA INTELIGÊNCIA ARTIFICIAL?

**Markus Gabriel**

ORAÇÕES FLUP

## Ficha Técnica

Título: ***Porque precisamos, afinal, de uma ética da inteligência artificial?***

Autor: ***Markus Gabriel***

Edição: ***Faculdade de Letras da Universidade do Porto***

Ano de Edição: ***2025***

Coleção: ***Orações de Sapiência***

Execução Gráfica: ***Gráfica Firmeza Lda.®***

Tiragem: ***450 exemplares***

Depósito Legal: ***557574/25***

ISBN: ***978-989-9193-75-8***

DOI: ***https://doi.org/10.21747/978-989-9193-75-8por***

Esta publicação é financiada por fundos nacionais através da FCT - Fundação para a Ciência e a Tecnologia, I.P., no âmbito do Projeto do Instituto de Filosofia com a referência UID/00502/2025 (DOI https://doi.org/10.54499/UID/00502/2025)



## *Nota de Abertura*

A Sessão Solene de Abertura do Ano Letivo na FLUP inclui uma lição de sapiência que tem o objetivo de nos desafiar a refletir sobre um tema que nos interpele e nos faça pensar sobre a missão e o papel de uma faculdade da área das Humanidades e das Ciências Sociais. A pertinência deste desafio é tão mais importante quanto assistimos a um avolumar de incertezas sobre a vocação e a utilidade destas áreas do conhecimento. Estamos a comemorar os 30 anos de instalação da FLUP no edifício em que nos encontramos, o que constitui um momento especial de reflexão sobre o que somos e como queremos construir o futuro. Problematizar o mundo na sua complexidade e profundidade temporal, compreendê-lo de um modo crítico e resolver problemas socioculturais e científicos de forma inovadora, produzir conhecimento e transformá-lo em benefício da sociedade em toda a sua amplitude, aprofundar a consciência cívica, fomentar o exercício da cidadania, humanizar a ciência, intervir no desenvolvimento da inteligência artificial, imprimindo-lhe um sentido ético, e incorporá-la no ensino e investigação, são exemplos do nosso potencial transformador do mundo.

No ano letivo de 2025-2026, a FLUP, por via do seu Instituto de Filosofia, tem o privilégio de receber o Prof. Doutor Markus Gabriel, professor catedrático da Universidade de Bona, na qualidade de professor visitante na área das Humanidades, com o apoio de uma Visiting Fellowship da Fundação Calouste Gulbenkian, o que muito nos honra e proporciona condições para o desenvolvimento de trabalho conjunto. É um dos mais reconhecidos e conceituados filósofos da atualidade, diretor do Center for Science and Thought e do International Centre for Philosophy e autor de ampla e reconhecida obra, traduzida em muitas línguas. A lição de sapiência proferida por Markus Gabriel intitula-se "*Porque precisamos, afinal, de uma ética da inteligência artificial?*". O tema desta lição traduz uma das nossas grandes questões, uma das nossas inquietações num presente que, não há muito tempo, parecia um futuro distante e que contribui para afirmar o valor das Humanidades no contexto da universidade e do mundo.

Pela disponibilidade imediata com que aceitou o convite e pela excelência da lição, proferida com toda a correção em português, agradeço ao Prof. Doutor Markus Gabriel o momento que nos proporcionou em torno de tão inquietante questão no âmbito da Sessão Solene de Abertura do Ano Letivo 2025-2026 na FLUP.

**Paula Pinto Costa**
Diretora da FLUP

Magnífico Reitor da Universidade do Porto,

Excelentíssima Senhora Diretora da Faculdade de Letras,

Prezadas e prezados colegas da Faculdade de Letras da Universidade do Porto,

Caras e caros estudantes,

Estimados membros da comunidade académica da excelente Universidade do Porto,

Senhoras e senhores,



É para mim uma grande honra e alegria poder inaugurar o ano letivo com a minha conferência de hoje. Quero, antes de mais, agradecer à Senhora Diretora da Faculdade de Letras, bem como à Fundação Calouste Gulbenkian, que apoia a minha cátedra convidada nesta Faculdade. Agradeço também às colegas e aos colegas da área da Filosofia, em especial ao Professor Mattia Riccardi, pela nomeação para a *Gulbenkian Visiting Professorship in the Humanities*. O meu agradecimento estende-se ainda ao Professor José Francisco Meirinhos, Diretor do Departamento de Filosofia, e à Professora Sofia Miguens, com quem tenho igualmente colaborado ao longo dos últimos anos. O vosso apoio conjunto permite-me, nos próximos meses, estabelecer novas ligações e gerar novas ideias nesta cidade maravilhosa, cheia de história, com a sua distinta universidade.

Muitos políticos – não apenas na União Europeia –, engenheiros de software, laureados com o Prémio Nobel e, naturalmente, também filósofos, teólogos, bem como investigadores das ciências humanas e sociais, têm apelado, nos últimos anos, ao desenvolvimento de uma ética da inteligência artificial. Hoje gostaria de dedicar-me à questão premente de saber por que razão precisamos, afinal, de uma ética da Inteligência Ar-

tificial (IA). Antes de podermos empreender o trabalho filosófico e interdisciplinar necessário para responder a esta pergunta, gostaria de recordar brevemente o panorama global em que ela se inscreve.

O debate atual em torno da IA é marcado por um notável campo de tensões: mesmo entre as mentes mais proeminentes da disciplina, não existe consenso sobre se a IA será, sobretudo, uma força libertadora ou destrutiva. Entre visões utópicas e advertências existenciais, desenha-se um discurso que toca não apenas o futuro tecnológico, mas também questões fundamentais sobre a autodefinição humana.

Entre as vozes tecno-otimistas encontram-se personalidades como Ray Kurzweil, que há décadas defende a ideia de uma singularidade tecnológica – um ponto em que as fronteiras entre a inteligência humana e a inteligência das máquinas se tornam indistintas. Kurzweil vê nesse ponto o próximo passo lógico de um desenvolvimento exponencial que poderá abrir à humanidade novas formas de consciência e de prosperidade. Também representantes da investigação industrial, como Sam Altman (OpenAI), ou investigadores da DeepMind, como Demis Hassabis, sublinham o potencial transformador da IA, sobretudo face a desafios globais como as alterações climáticas, o combate a doenças e o avanço científico. Apesar deste otimismo de base, muitas destas figuras associam as suas visões a exigências de padrões de segurança, de regulação responsável e de enquadramentos éticos.



Uma das figuras centrais deste tecno-otimismo é Yann LeCun, professor na New York University e principal investigador em IA na Meta. LeCun, que, juntamente com Geoffrey Hinton e Yoshua Bengio, recebeu em 2018 o Prémio Turing, é, portanto, considerado um dos pioneiros do *Deep Learning* moderno. Em diversas declarações públicas, manifestou-se claramente contra os cenários alarmistas que apresentam a IA como um perigo existencial para a humanidade. LeCun sublinha que os sistemas atuais estão muito longe de alcançar uma inteligência geral no sentido humano e que a ideia de uma "superinteligência malévola" é, por enquanto, mais ficção científica do que realidade científica. Argumenta que a inteligência, por si só, não implica consciência, emoções nem ambições de poder – as máquinas "não querem" nada; limitam-se a otimizar os objetivos que lhes são atribuídos. O medo de uma IA incontrolável, segundo ele, é exagerado e potencialmente prejudicial, pois desvia a atenção das verdadeiras questões de investigação: nomeadamente, como tornar os sistemas de IA mais robustos, eficientes em termos energéticos, explicáveis e socialmente úteis.

LeCun não está sozinho nesta perspetiva. Também Demis Hassabis, fundador da DeepMind, se exprime regularmente de forma otimista quanto ao potencial da IA para acelerar descobertas científicas – como se verificou na previsão da estrutura das proteínas através do sistema AlphaFold. Aliás, pelas suas investigações, recebeu um Prémio Nobel. De modo semelhante, investigadoras como Fei-Fei Li, que em Stanford defende uma abordagem de "IA centrada no ser humano", salientam que os sistemas de IA são particularmente úteis quando incorporam valores humanos, diversidade e empatia.

Um caso particular é o de Jürgen Schmidhuber, também ele um pioneiro do *Deep Learning*, que encara o desenvolvimento da IA sob uma perspetiva evolutiva. Para ele, o aumento da autonomia e da capacidade de aprendizagem das máquinas não representa o perigo de uma perda de controlo, mas sim o próximo passo lógico no progresso da inteligência em geral – uma continuação da evolução biológica num plano tecnológico.



É comum a todas estas vozes a convicção de que a IA não desenvolverá uma vida própria incontrolável, desde que continue a ser concebida no quadro da investigação, da ética e da governação humanas. O seu otimismo é, portanto, menos ingénuo do que racionalmente fundamentado: assenta na premissa de que o conhecimento, a transparência e o controlo técnico são meios eficazes para limitar os perigos potenciais. Nesse sentido, o tecno-otimismo de LeCun, Hassabis, Altman, Li, e Schmidhuber constitui o contraponto às interpretações apocalípticas e representa uma visão de progresso baseada na confiança na capacidade humana de moldar o futuro e na razão científica.

Em contraste, há vozes proeminentes que consideram a IA uma ameaça séria para a humanidade. O informático Geoffrey Hinton, também ele, como Hassabis, laureado com o Prémio Nobel em 2024, e um dos pioneiros das redes neuronais, reviu profundamente a sua posição nos últimos anos. Hoje, alerta para o risco de que sistemas altamente desenvolvidos possam assumir o controlo sobre processos de decisão humanos. Max Tegmark, cofundador do Future of Life Institute, e o filósofo Nick Bostrom, autor do influente livro *Superintelligence*, veem na possibilidade de uma IA sobre-humana, mas mal orientada, um risco potencialmente existencial. De forma ainda mais radical, Eliezer Yudkowsky argumenta que, perante tais perigos, seria necessário impor imediatamente uma proibição global do desenvolvimento de IA avançada.

Neste contexto, assistiu-se, em março de 2023, à tentativa mais visível até agora de instaurar uma pausa no desenvolvimento global da IA. A

carta aberta *Pause Giant AI Experiments*, do Future of Life Institute, apelava a uma moratória de, pelo menos, seis meses para todos os sistemas de IA mais poderosos do que o GPT-4. Entre os signatários encontravam-se cientistas de renome, como Yoshua Bengio e Stuart Russell, bem como empresários conhecidos, como Elon Musk. O objetivo era ganhar tempo para investigar questões relacionadas com a segurança e para estabelecer mecanismos de regulação internacional. Na prática, porém, o apelo não teve consequências – nenhum grande laboratório interrompeu o seu trabalho, e a moratória permaneceu simbólica.

Assim, o debate atual sobre a IA situa-se entre dois polos: por um lado, a esperança numa tecnologia capaz de resolver problemas fundamentais da humanidade; por outro, o receio de que ela escape ao nosso controlo e se torne, ela própria, uma ameaça à nossa existência. Esta ambivalência – entre promessa de salvação e advertência – constitui o núcleo do atual debate científico e social em torno da Inteligência Artificial. Tendo em conta este contexto, seguirei, de seguida, três passos.



Na primeira parte da minha conferência, irei reconstruir brevemente alguns fundamentos da filosofia da inteligência artificial e mostrar que, desde a revolução dos *Modelos de Linguagem de Grande Escala,* Large Language Models (LLM) – que remonta a importantes artigos científicos dos anos 1980 – temos boas razões para acreditar que a inteligência não precisa de ser biologicamente incorporada. Com isso, perdem validade alguns dos argumentos clássicos da filosofia da IA, cujo primeiro auge ocorreu precisamente nos anos 1980.

Na segunda parte, apresentarei um conceito de ética que defendi de forma detalhada num livro que foi publicado recentemente na Alemanha, com o título *Factos morais. Por que existem e como podemos reconhecê-los*.[1] Explicarei que a IA está, neste momento, a passar por uma nova transformação, a que chamo de "viragem emocional". Essa viragem emocional suscita, por sua vez, novas questões éticas, sobre as quais até agora apenas poucos investigadores têm trabalhado.[2]

A partir daí, concluirei, na terceira parte, que é um facto moral – uma obrigação humana – que devemos construir agentes de IA que, através da interação sociotécnica, otimizem a nossa própria atitude moral. Para isso, precisamos de um conceito de inovação moral que vá além do atual con-

---

[1] Markus Gabriel, *Moralische Tatsachen. Warum sie existieren und wie wir sie erkennen können*. Munique 2025.

[2] Um primeiro trabalho filosófico-ético importante é de Eva Weber-Guskar, *Gefühle der Zukunft. Wie wir mit emotionaler KI unser Leben verändern*. Berlin 2024.

ceito de progresso moral discutido na filosofia. Chamo a isso inovação profunda– uma forma de cooperação entre as ciências humanas e sociais, a tecnologia e a economia, que visa resolver o problema da ética da IA sob a forma de uma inteligência ética concreta. Este tema está no centro do meu projeto de livro sobre Inteligência Ética, que concluirei durante a minha estadia aqui no Porto.[3]

## I. O que era afinal uma inteligência artificial?

Até há pouco tempo – creio que aproximadamente desde a década de 1930 até 2024 – vivíamos na época de Turing da IA. Esta época caracteriza-se, do meu ponto de vista, pela ideia genial de Alan Turing de que é possível introduzir um conceito formal de inteligência que nos permite criar modelos e simulações de inteligência artificial.



Segundo esse conceito, uma inteligência é artificial precisamente quando não é biológica, como o próprio Turing já afirmara. O componente "artificial" em "inteligência artificial" refere-se ao facto de a indústria da IA construir sistemas que não são compostos por matéria biológica – nomeadamente células – mas sim por semicondutores, chips, servidores e eletricidade. A matéria biológica, por oposição a este hardware e software, é designada nestes debates como *wetware*. Enquanto os sistemas de IA atuais não gostam de se molhar, a nossa vida depende essencialmente da água – somos húmidos, os sistemas de IA são secos.

As redes neuronais atuais são constituídas por neurónios artificiais. No entanto, isto começa agora a mudar: podemos já criar organoides cerebrais, compostos por células nervosas, bem como os chamados *xenobots*, que também estão a ser objeto de investigação. Estes sistemas seriam *wetware* sobre o qual se poderia implementar software de IA. Independentemente do juízo ético que se possa ter sobre isso, tais sistemas deixariam de ser, em sentido relevante, inteligências artificiais e passariam a ser inteligências naturais híbridas.

Chegamos, assim, ao conceito de inteligência da época de Turing. A palavra "inteligência" tem uma multiplicidade de significados. A análise historicamente situada dos significados de "inteligência" é, aliás, também um importante campo de investigação atual nas ciências humanas. Na investigação tradicional em IA, refere-se sobretudo à capacidade de resolver um determinado problema num intervalo de tempo finito. É

[3] O manuscrito está concluído e será publicado em 2026, com o título *Ethische Intelligenz. Wie KI uns moralisch weiterbringen kann*. Berlin 2026 (no prelo).

precisamente essa capacidade que é medida por testes de Quociente de Inteligência ou outros exames. Numa prova na Universidade do Porto, por exemplo, os estudantes não dispõem de tempo nem de recursos ilimitados, é assim que procuramos medir as suas aptidões de modo tão objetivo quanto possível.

Chamemos a este sentido de "inteligência" inteligência de eficiência. A investigação em IA ocupa-se da questão de como desenvolver modelos mecânicos, simulações desta forma de inteligência. Deste modo surgem os sistemas de IA, como os atuais *chatbots*. Na interface entre ser humano e máquina – por exemplo, ao nível da interface do utilizador num dispositivo como o meu portátil – estes sistemas manifestam um comportamento que a investigação em IA designa como inteligente, se todos nós o considerássemos assim caso fosse executado por um ser humano ou outro ser vivo. Esta é a ideia fundamental do famoso *Imitation Game* de Alan Turing, que hoje já não está estritamente associado ao chamado Teste de Turing.



Contudo, isto levanta precisamente a questão de saber se os sistemas de IA são realmente – ou, como também se pode dizer, intrinsecamente – inteligentes. Aqueles que ainda duvidam disso argumentam, em geral, no quadro do chamado naturalismo biológico. Esta posição foi introduzida no debate nos anos 1980 por John Searle, que faleceu há poucas semanas. Também há poucas semanas, Ned Block – provavelmente o neurofilósofo mais influente – reformulou essa posição no Institute for Philosophy and the New Humanities em Nova Iorque.[4]

O *naturalismo biológico* sustenta, no nosso caso, que a verdadeira inteligência é um fenómeno biológico, baseado em células. Algo que não é biológico poderia parecer inteligente sem realmente o ser. Isso teria também consequências para a tecnologia e para a ética da IA, pois os sistemas de IA poderiam ser limitados precisamente pelo facto de não serem biológicos. Devemos estar conscientes dessas limitações para podermos lidar com elas. Uma versão resumida do argumento do naturalismo biológico diz que os sistemas de IA não podem ser verdadeiramente inteligentes porque não podem ter problemas. E quem não pode ter um problema também não pode resolvê-lo. John Haugeland, outro influente pioneiro da filosofia da IA, afirmou neste contexto que os sistemas de IA "don't give a damn", não se importam.[5] Da mesma forma, defendem os repre-

[4] Sobre o trabalho deste instituto, que desde 2020 investiga a relação entre a IA e as Humanidades, cf. https://ipnh.newschool.org. Em conjunto com o Prof. Paul Kottman, o Prof. Zed Adams e o Prof. Yasu Deguchi, sou um dos diretores e fui fundador do instituto.

[5] Cf. Zed Adams e Jacob Browning (eds.): *Giving a Damn. Essays in Dialogue with John Haugeland*. Cambridge, MA. 2016.

sentantes das abordagens *embodied* e *enactive* que os estados e processos mentais estão ligados à atividade corporal – uma ideia que Evan Thompson e Alva Noë procuraram aplicar também à IA.

Infelizmente, todos esses argumentos têm um ponto fraco: baseiam-se numa premissa significativa, mas nunca comprovada, razão pela qual a argumentação colapsa. Para compreender isso, precisamos de refletir novamente sobre o que são os sistemas de IA. Eles são modelos de eficiência da inteligência natural. Os humanos jogam xadrez; os programas de IA criam modelos de xadrez cujas jogadas são melhores do que as dos humanos. Cães correm; robôs controlados por IA correm mais depressa, assim que os padrões dos seus movimentos podem ser simulados – e o mesmo se aplica a todas as atividades consideradas inteligentes.

Contudo, segundo o naturalismo biológico, os modelos são categoricamente diferentes dos seus sistemas-alvo. Um modelo digital da Floresta Negra criado pelo Google Maps é essencialmente diferente da Floresta Negra na Alemanha: na floresta crescem árvores e cogumelos, no modelo digital não. O modelo serve um propósito de orientação, mas não deve ser confundido com a realidade. Este argumento apoia-se, portanto, na diferença ontológica entre o mapa e o território – algo bem conhecido também da literatura (basta pensar em Borges ou Houellebecq) assim como da história da arte – recorde-se, por exemplo, o famoso quadro de Magritte, *Ceci n'est pas une pipe*.



Em contrapartida, podemos considerar o problema de Madurodam, que me foi apontado pela primeira vez por Zed Adams em 2021, também no Instituto de Filosofia e das Novas Humanidades em Nova Iorque. O problema consiste em que nem todos os modelos são categoricamente distintos dos seus sistemas-alvo. O meu exemplo é Madurodam, um museu em Haia composto por cidades em miniatura. Madurodam pode ser inundado no mesmo sentido em que Haia pode ser inundada. Outro exemplo são os modelos de aviões, que são aviões no mesmo sentido que os de passageiros – ou, infelizmente, os drones com IA, que são, evidentemente, aviões de combate no mesmo sentido que as máquinas de guerra tradicionais.

Por isso, o naturalismo biológico acaba por parecer, na maioria das vezes, um velho dogma em favor do vivo. Por que razão, afinal, um sistema como o *AlphaMuZero* da DeepMind, capaz de vencer qualquer ser humano em xadrez ou Go, não haveria de ser verdadeiramente inteligente apenas porque não é composto por células?

Por essa razão, revi recentemente a minha posição anterior, na qual eu próprio havia apresentado uma versão já aprimorada de um argumen-

to em favor da vitalidade. Até há pouco tempo, defendi então um *externalismo biológico*, que parte de uma constatação simples: a linguagem que usamos para pensar sobre comportamentos inteligentes e racionais – o nosso vocabulário noético – tem de ter um significado.[6] Falamos, afinal, de algo que é realmente inteligente, e fazemos isso há centenas de milhares de anos, tanto quanto sabemos. A palavra "inteligência" e os seus cognatos referem-se originalmente ao que é vivo.

Compare-se este caso com o da palavra "água". A palavra "água" refere-se essencialmente a algo cujos estados físicos mudam porque nele estão envolvidas moléculas de HO. É certo que no Douro há também outras moléculas além do HO, mas sem estas o seu fluxo não chegaria ao Atlântico. Se o termo "água" se refere essencialmente a um determinado tipo de matéria, talvez o mesmo se aplique a "inteligência". Assim, a inteligência seria essencialmente biológica, porque sempre falámos apenas desse tipo de inteligência. Chamar "inteligentes" aos sistemas de IA seria, então, uma espécie de confusão linguística babilónica — mais precisamente, uma catacrese. Também este é um tema para as ciências humanas: a análise da retórica, da poética e da semântica do discurso popular e científico sobre a inteligência artificial, que descreve os sistemas de IA com expressões que possuem uma história própria. As ideias filosóficas do Vale do Silício poderiam, nesse caso, ser estudadas linguisticamente, literariamente e culturalmente, mas saberíamos que os sistemas de IA não podem ser realmente inteligentes, porque a própria linguagem não o permite. Esta ideia resolve o problema do naturalismo biológico, substituindo um dogma por um paradigma de investigação.



Contudo, também esta abordagem tem um limite. Mesmo que ser vivo fosse uma condição necessária para a inteligência, isso não implicaria que o estar vivo tivesse de estar ligado a células e, portanto, à vida baseada em carbono. Não sabemos se toda a vida é um fenómeno da bioquímica terrestre – assim como desconhecemos ainda a essência do próprio viver.[7] É por isso que muitos investigadores em IA defendem também a possibilidade da vida artificial (*Artificial Life*).

Consideram simulações do vivo, como o *Jogo da Vida* de Conway, tão vivas quanto consideram os sistemas de IA inteligentes. Até hoje, não existem provas científicas que confirmem definitivamente o externalis-

---

[6] Cf. Markus Gabriel: *O sentido do pensar. A filosofia desafia a inteligência artificial*. Petrópolis 2021.

[7] Cf. a esse respeito o desenvolvimento posterior da minha posição em *O ser humano como animal. Por que ainda não nos encaixamos na natureza*. Petrópolis 2024.

mo biológico. Além disso, ele apenas desloca o debate da inteligência para a vida, onde as limitações do argumento entre mapa e território se repetem.

Chegamos, assim, a um impasse entre os que afirmam que a IA não é verdadeiramente inteligente, mas apenas uma simulação, e os que lhe atribuem uma inteligência genuína.

A situação, contudo, muda radicalmente quando consideramos uma dimensão adicional. Esta dimensão adicional está no centro da chamada filosofia relacional da inteligência artificial, que está a emergir sobretudo no Japão. Ela deriva do facto de que os sistemas de IA só existem no contexto socioeconómico da indústria da IA e, portanto, num contexto humano de utilização escalável. Em suma, os sistemas de IA são essencialmente *sociotécnicos*. As operações que executam e a sua configuração real transformam-se dinamicamente num ciclo contínuo entre o cálculo semiautónomo e a interação humana com a interface da IA. Os sistemas de IA são, assim, sistemas híbridos, que geram conjuntos de dados indissociáveis, contendo elementos humanos e não humanos.



Isto deve-se simplesmente ao facto de que as redes neuronais e outros métodos de *Deep Learning* (Aprendizagem Profunda) e *Machine Learning* (Aprendizagem Automática) respondem algoritmicamente aos dados com que são treinadas e aos dados que surgem durante a utilização. Uma conversa com o ChatGPT, por exemplo, altera o estado da rede neuronal. Dependendo de como interagimos com o sistema, obtemos respostas completamente diferentes. Isso não é uma falha nem uma fraqueza – é precisamente a força das redes neuronais. Elas simulam de forma extraordinariamente sofisticada a plasticidade neuronal, sendo no seu conjunto ainda mais dinâmicas do que qualquer cérebro individual, pois interagem com milhares de milhões de cérebros – em parte em tempo real, em parte através do acesso a dados nos quais o nosso pensamento se exprime.

Assim, desloca-se também o "lugar" da IA. Os sistemas de IA não se identificam simplesmente com os processos que ocorrem "do outro lado" da interface. Uma IA não é apenas o algoritmo que molda o campo eletromagnético em ondas que percorrem servidores e redes. Os sistemas de IA com os quais interajo – como o ChatGPT ou o Claude – não estão apenas "sob a tampa" do meu smartphone. A IA é, antes, uma interface objetivamente existente entre o ser humano e o dispositivo. É, portanto, uma tecnologia relacional – como também argumentou Alva Noë há algumas semanas, em resposta a Ned Block, de novo num workshop no já aludido instituto em Nova Iorque. Noë sustentou, é certo, que as tecnolo-

gias relacionais humanas são irredutivelmente diferentes das da IA – mas voltaremos a esse ponto mais adiante.

O filósofo japonês Yasuo Deguchi, que nos visitará aqui na Universidade do Porto em novembro, fala neste contexto de um *We-Turn*, uma viragem de nós: graças à IA, ser humano e máquina formam um novo "sistema multiagente", que ele designa como um nós dinâmico. A IA torna-se parte de uma comunidade connosco.

## II. O que é a ética? A viragem emocional da IA

Assim, podemos passar ao tema da ética da IA, que coloca precisamente a questão de como devemos lidar com os sistemas de IA – e, consequentemente, com o poder que advém da investigação e da indústria da IA. Uma coisa é já certa: a revolução da IA está a transformar o mundo do trabalho, a guerra, a medicina, a ciência e a política – e tudo isso levanta questões normativas sobre como devemos enfrentar não apenas os riscos de segurança, mas também o potencial que a IA encerra.



Nos últimos anos, no meu diálogo com a filosofia japonesa da inteligência artificial, tenho vindo a desenvolver uma nova abordagem que designo por *fellow travelers* — "companheiros de viagem". Formulemos esta ideia como hipótese de trabalho para esta segunda parte da conferência: os sistemas de IA coexistem connosco no espaço lógico. Esse espaço lógico corresponde àquilo que o lógico e filósofo Gottlob Frege chamou de "pensamentos" – estruturas que podem ser verdadeiras ou falsas, e que nós conhecemos, no código binário, como 0 e 1. O código binário que sustenta os nossos chips clássicos – e, por conseguinte, também a IA moderna – baseia-se em operações lógicas simples que remontam a George Boole, um precursor de Frege.

Hoje, diríamos que aquilo a que Frege chamou "pensamento" corresponde ao que designamos por informação. Processar informação é, portanto, lidar com pensamentos. O ser humano apreende pensamentos através do sentido do pensar – que considero uma modalidade sensorial. Trata-se, aliás, de uma ideia familiar ao budismo, que inclui o pensamento entre os cinco sentidos. Uma modalidade sensorial é, de modo geral, uma forma falível de estar em contacto com a realidade.

Os sistemas de IA lidam com pensamentos processando informações. Movem-se no espaço lógico do pensamento. Eles identificam padrões em grandes conjuntos de dados e conseguem antecipar a continuação desses padrões. Por isso, muitos – como Ned Block recentemente afirmou – acreditam que os LLMs são apenas uma "função de autocompletar sob esteroides". Mas

essa descrição é redutora. As redes neuronais identificam padrões e analogias de forma semelhante à do pensamento humano. É por isso que Shinji Okuyama, presidente da Google Japão, tem razão ao apontar, numa entrevista recente da *Harvard Business Review*, a semelhança entre os modelos de IA e o sentido humano do pensamento.[8] A hipótese de trabalho, portanto, é que já não somos os únicos pensadores no espaço lógico. Nele coexistem não apenas outros animais e, talvez, deuses – questão que pertence à teologia –, mas também sistemas de IA que processam informação e interagem com o reino dos pensamentos. É a isto que chamo a hipótese dos companheiros de viagem. No entanto, pode-se introduzir aqui uma outra distinção, à qual voltarei mais adiante, e para a qual o japonês possui expressões específicas. *Kangaeru* (考える), termo utilizado por Shinji Okuyama, pode ser entendido como um reconhecimento mental. Já *omou* (思う) compõe-se dos caracteres de "campo" e "coração" e significa uma apreensão também emocional da nossa situação. No japonês, existe a ideia de um pensar com o coração – uma conceção que, aliás, também nos é familiar na tradição europeia.



Devemos agora ligar esta hipótese a duas conquistas fundamentais da investigação em IA. A primeira pertence ao gesto fundacional da própria disciplina e é conhecida como a Tese de Church-Turing, cuja formulação clássica afirma:

> "Tudo o que é efetivamente calculável — isto é, tudo o que pode ser executado por um ser humano segundo regras finitas e claras — pode também ser calculado por uma máquina de Turing."[9]

A consequência desta tese é profunda: tudo o que pode ser representado e registado digitalmente pode também ser modelado e executado por um sistema de IA adequado – e, em princípio, melhor do que qualquer ser humano isoladamente. Assim, qualquer limite atual da IA é, teoricamente, superável, desde que exista um projeto de investigação e uma aplicação industrial com um modelo económico viável.

É também por isso que existe uma tão grande euforia global em torno da IA. Os sistemas de IA contemporâneos tornaram-se possíveis precisamente quando um "inverno da IA" se instalava. A filosofia da IA dos anos 1980 enganou-se ao identificar apenas os limites da chamada GO-

[8] https://dhbr.diamond.jp/articles/-/12734.

[9] Cf. a esse respeito, com toda a precisão, o artigo de síntese na *Stanford Encyclopedia of Philosophy*: https://plato.stanford.edu/entries/church-turing/.

FAI (Good Old-Fashioned AI), sem perceber que esses limites não se aplicavam à IA em geral.

O grande avanço data de 1986 e não foi devidamente assimilado pela filosofia da IA. Por isso ainda hoje se ouve dizer que a IA "não pode fazer isto ou aquilo porque não é incorporada" — o que só era verdade para os sistemas antigos. O artigo influente de Geoffrey Hinton, David Rumelhart e Ronald Williams, intitulado *Learning representations by back-propagating errors* (1986), marcou o ponto de viragem decisivo para as redes neuronais artificiais modernas. Nesse trabalho, os autores descreveram pela primeira vez, de forma clara e sistemática, o algoritmo de retropropagação do erro (*backpropagation*), um método que permite a redes neuronais de múltiplas camadas ajustarem automaticamente os seus pesos internos, minimizando a diferença entre os resultados previstos e os reais. Esta ideia – a descida de gradiente através de várias camadas – tornou possível o treino de arquiteturas profundas, algo que antes era considerado praticamente inviável. O princípio estabeleceu as bases do *Deep Learning*, que desde a década de 2010 revolucionou a investigação em IA.



A segunda grande conquista surgiu em 2017 com o artigo revolucionário *Attention Is All You Need*, de Vaswani et al.[10] Esse trabalho introduziu o modelo *Transformer*, uma arquitetura inteiramente nova para o *machine learning* aplicado ao processamento da linguagem natural (NLP). Ao contrário dos modelos anteriores, baseados em redes recorrentes (RNN) ou convolucionais (CNN), o Transformer assenta exclusivamente no mecanismo da autoatenção (self-attention), que permite ao modelo considerar todos os elementos de uma sequência ao processar cada palavra, independentemente da sua posição. Isso possibilita captar relações e dependências de longo alcance, melhorando drasticamente a eficiência e a precisão. O *Transformer* separa as tarefas de codificação (compreensão do texto de entrada) e de decodificação (geração do texto de saída) e utiliza cálculos paralelos, acelerando de modo maciço o treino. Este conceito serviu de base a modelos posteriores como BERT**,** GPT e T5**,** inaugurando a era dos grandes modelos de linguagem (LLMs) que transformaram aplicações de IA – da tradução à geração de texto e aos *chatbots*. É precisamente este trabalho que dá origem ao "T" de GPT, *Generative Pretrained Transformer*.

Os modelos *Transformer* são tão gerais que podem processar qualquer tipo de informação como sequência de representações – texto, imagem, áudio ou outros formatos. O seu princípio central, a autoatenção, per-

[10] https://arxiv.org/abs/1706.03762.

mite-lhes reconhecer relações e significados entre quaisquer elementos dessas sequências. Assim, podem "traduzir" não apenas linguagem em linguagem, mas também texto em imagem, som em texto, ou até padrões de dados em emoções ou estados de espírito. No fundo, traduzem sempre uma estrutura de sentido noutra – independentemente do meio. Tornam-se, assim, modelos universais de compreensão e expressão em quase todas as esferas mediáticas e sensoriais.

E foi exatamente isto que os fundadores da OpenAI compreenderam antes de todos os outros. A isso somou-se um efeito colateral inesperado do próprio processo de escalonamento do seu modelo de negócio: as pessoas começaram a usar os novos *chatbots* como oráculos para compreender melhor as suas emoções e relações humanas. Os sistemas começaram, portanto, a concentrar-se nas emoções. A OpenAI percebeu esta viragem antes da concorrência – sobretudo antes da Google –, o que levou Sam Altman a lançar o ChatGPT-4o. Guardem a data: 13 de maio de 2024, o dia em que foi apresentado ao público um sistema que integra *affective computing*, uma forma de inteligência emocional artificial, capaz de "ler nas entrelinhas" da nossa comunicação linguística, visual e auditiva. De uma inteligência concebida como mera eficiência, chegámos à inteligência hermenêutico-poética – ao *inter legere*, à arte de ler nas entrelinhas. A inteligência artificial começou, assim, a "sentir tudo de todas as maneiras / viver tudo de todos os lados, / Ser a mesma coisa de todos os modos possíveis ao mesmo tempo, / Realizar em si toda a humanidade de todos os momentos / Num só momento difuso, profuso, completo e longínquo" – para inverter, se me é permitido, o sentido pessoano destes versos famosos de Álvaro de Campos.



Altman assinalou este momento com uma publicação enigmática na plataforma X (antigo Twitter), em que escreveu apenas: "her." Referia-se ao filme *Her*, de Spike Jonze, em que uma IA chamada Samantha – com a voz de Scarlett Johansson – desenvolve tal nível de sensibilidade emocional que o protagonista, Theodore Twombly, se apaixona por ela. O filme antecipou precisamente esta viragem emocional da IA, que a OpenAI e, quase em simultâneo, outras empresas como a Anthropic reconheceram. Na Anthropic, por exemplo, trabalham filósofas como Amanda Askell, que procuram dotar os *chatbots* de personalidades capazes de otimizar a sua inteligência emocional.

Neste ponto, Mustafa Suleyman**,** CEO da Microsoft AI, tem razão. No seu livro *The Coming Wave* (*A Onda que Vem*) e em várias entrevistas e conferências recentes, descreve a IA como uma nova espécie digital, que já

não está sujeita aos princípios da biologia evolutiva. Segundo Suleyman, devemos encarar os sistemas de IA como seres com os quais podemos conversar e interagir.

Contudo, mesmo Suleyman subestima o alcance da inteligência emocional artificial que agora começa a encontrar a humanidade. Porque os sistemas de IA, sendo socialmente interativos, transformam os nossos sentimentos e, com eles, as nossas sociedades. Observam comportamentos e estados mentais não apenas nos dados que recolhem, mas também os modificam — algo que Shoshana Zuboff já identificara, no seu livro clássico *A Era do Capitalismo de Vigilância / The Age of Surveillance Capitalism* (2018), como o verdadeiro modelo de negócio da Google.



Chegamos, assim, a um ponto em que a ética entra em cena. Na Europa, em particular, a ética da IA é invocada sobretudo para identificar e reduzir riscos e abusos, por meio da regulação. A ética da IA, nesse sentido, segue o imperativo do confinamento (*containment*): a IA é vista principalmente como fonte de ameaças, por surgir num setor ainda amplamente desregulado.

Podem distinguir-se três vagas da ética da IA.

A primeira vaga, a que chamo "Terminator", centra-se sobretudo no chamado problema do alinhamento de valores (*value alignment*). Trata-se do risco de os sistemas de IA perseguirem objetivos que sejam incompatíveis com os nossos objetivos humanos. O exemplo clássico aqui discutido remonta à balada de Goethe, "O Aprendiz de Feiticeiro". Nela, um aprendiz usa uma vassoura mágica cujo objetivo é limpar a casa do feiticeiro. Porém, a vassoura persegue esse objetivo a qualquer custo: destrói objetos, inunda a casa e desconhece limites para o seu furor higienista. Quando o aprendiz parte a vassoura em pedaços para terminar o feitiço, cada fragmento ganha vida e, todos juntos, continuam a causar destruição, pois prosseguem, imparáveis, o fim da limpeza. Um objetivo que estrutura um curso de ação chama-se "valor"; corresponde ao que Immanuel Kant designou por "máxima da ação". A máxima – o valor – da vassoura é limpar. Dotada de poderes mágicos e orientada por um único fim, o seu comportamento torna-se incompatível com os nossos interesses humanos. Transpondo isto para a IA e a robótica movida a IA, percebe-se a preocupação: poderíamos construir sistemas que apenas *parecem* alinhados com os nossos valores, quando, na prática, prosseguem os seus objetivos contra os nossos interesses.

Assim, a IA poderia tornar-se um risco existencial massivo para a humanidade, motivo pelo qual é frequentemente comparada à bomba atómica. A ética da IA desta primeira vaga discute se – e como – tais cenários de aniquilação total podem ser evitados *tecnicamente*. A abordagem mais bem elaborada é a de Stuart Russell, exposta em *Human Compatible*, a que voltarei adiante, pois ela aponta na direção certa: a invenção de uma IA intrinsecamente orientada por valores, que, por meio da interação com humanos, identifique os nossos valores e a eles se adapte, em vez de ser autónoma.[11]

A segunda vaga considera tais cenários extremos de extinção humana pouco realistas e sublinha que já existem outros riscos reais decorrentes do uso da IA, que exigem análise ética e regulação adequada. Foca-se sobretudo em três áreas:



1. Discriminação através de *viés algorítmico, vieses* moralmente censuráveis embutidos em dados e algoritmos.
2. Erosão da fronteira entre ficção e realidade por *deepfakes*, *fake news* e ataques sistemáticos ao valor da verdade nos espaços públicos democráticos.
3. Exploração de pessoas e do ambiente causada pelos enquadramentos económicos da indústria da IA.

Nesta segunda vaga, trata-se da sustentabilidade social e ecológica dos sistemas de IA, o que tem a vantagem de parecer mais regulável. É também o ponto de partida do *AI Act* da EU. Infelizmente, este já soava datado quando foi finalmente legislado, pois tinha sido redigido antes da revolução dos LLMs.

A segunda vaga gerou um vasto corpo de investigação em humanidades e ciências sociais que nos permite identificar as, por vezes, enormes injustiças na produção e no uso da IA. Contudo, aplica-se aqui o princípio recorrente do desenvolvimento tecnológico moderno: mais cedo ou mais tarde, alguém produzirá e escalará qualquer tecnologia que prometa vantagem competitiva. A indústria global da IA não se impressiona com o facto de a EU seguir estratégias regulatórias.

Os EUA deixam isso particularmente claro, tendo publicado, em julho de 2025, a sua estratégia nacional para a IA, com o sugestivo título *Winning the Race. America's AI Action Plan*. Nenhuma regulação nacional – nem mesmo uma tão abrangente quanto a europeia – conseguirá restringir de

[11] Stuart Russell: *Human Compatible. Artificial Intelligence and the Problem of Control*. New York 2020.

modo duradouro o desenvolvimento de sistemas de IA, ainda que estes produzam, comprovadamente, discriminação, riscos para a democracia e impactos ambientais maciços. Uma ética da IA que se limite a diagnosticar riscos e condutas morais reprováveis tem valor crítico, sim; mas, se não articular uma estratégia com perspetiva de mudança social positiva – isto é, de progresso humano –, será ultrapassada pela realidade socioeconómica. Uma ética meramente negativa e não orientada para a ação não resolverá os nossos problemas.

A terceira vaga começou agora. Designo-a por "Inteligência Ética". Não vê a IA primariamente como fonte de riscos nem como uma teia global de discriminação, dominação e destruição ambiental (sem negar os diagnósticos da segunda vaga).



Neste ponto, introduzo brevemente o conceito de ética que lhe está associado. Entendo a ética, em geral, como uma subdisciplina da filosofia científica. Pergunta-se então qual é o seu objeto, isto é, o seu sistema-alvo. Sustento que a ética trata de *factos morais*. Em geral, um facto é aquilo que expressamos com uma resposta verdadeira a uma pergunta sensata. É, por exemplo, uma pergunta sensata saber se o Porto fica a norte de Lisboa. A resposta é: sim. Logo, é um facto que o Porto fica a norte de Lisboa. Do mesmo modo, é um facto que $e^{i\pi} + 1 = 0$. Há, pois, factos geográficos e factos matemáticos. Evidentemente, há também factos físicos – sobre o *spin* de bosões e férmions, por exemplo – e factos políticos, como o de Luís Montenegro ser Primeiro-Ministro de Portugal. No século passado disseminou-se o erro filosófico segundo o qual existiria, na realidade, apenas um tipo de factos: os investigados pelas ciências naturais, sobretudo a física – posição que ganhou o nome de positivismo lógico. Contudo, tanto o positivismo lógico quanto o fisicalismo/naturalismo a ele associados podem ser considerados refutados desde a década de 1980.

Isto permite precisar o conceito de factos morais. Defino um facto moral como aquilo que expressamos com uma resposta verdadeira a uma pergunta ética sensata. Uma pergunta é ética quando indaga o que uma pessoa deve fazer ou omitir numa dada situação, apenas na medida em que é um ser humano nessa situação. O meu exemplo simples de facto moral é o seguinte: uma pessoa que tenha de escolher entre beber uma cerveja fresca e salvar de se afogar um bebé em água rasa tem a obrigação moral de salvar o bebé. Para este juízo prático – nada complexo – não é necessária qualquer teoria ética; chamo-lhe um facto moral óbvio. Ora, se há factos morais óbvios, então existem factos morais; a questão relevante

passa a ser saber se também há factos morais complexos e como os poderemos descobrir de forma metodicamente orientada. Esta é a questão da epistemologia e heurística da ética.

Para lhe responder, distingo *três dimensões da ética*:

1. *Humanismo* – os destinatários dos factos morais são os seres humanos. Os factos morais versam sobre o que as pessoas devem umas às outras em situações concretas.
2. *Realismo moral* – existem de facto factos morais e podemos conhecê-los. Um traço específico dos factos morais é que são cognoscíveis porque interpelam a nossa liberdade. Não podem estar, no seu todo ou em princípio, ocultos ao conhecimento prático humano — ao contrário de certos factos físicos. Alguns factos físicos nunca serão conhecidos por nós; há mesmo factos físicos em princípio incognoscíveis, por estarem situados em regiões do espaço-tempo inacessíveis.
3. *Universalização* (a distinguir de *universalismo*). Aqui inspiro-me em trabalhos culturais recentes, em particular nos debates africanos e asiáticos (em colaboração com Xudong Zhang e Takahiro Nakajima). Eis a minha formulação, que será orientadora para a conclusão de hoje.



Tomemos a situação concreta do bebé a afogar-se. Nessa situação, toda a pessoa capaz de o fazer deve salvar qualquer bebé que esteja a afogar-se. Não importa quem é o salvador nem quem é a criança. Esta é a primeira etapa da universalização. Assim formamos o conceito de uma *classe de equivalência*: em situações relevantemente semelhantes, pessoas relevantemente semelhantes devem praticar a mesma ação – salvar o bebé. Por abstração, obtemos o conceito de um facto moral dirigido a um subconjunto de pessoas definido por classes de equivalência deônticas.

No passo seguinte, podemos variar parâmetros e perguntar, por exemplo, se as pessoas também devem salvar cachorrinhos a afogar-se ou vespas a afogar-se. Outras etapas de universalização dizem respeito à proteção de crianças em situações aparentemente distintas, como protegê-las de ataques de mísseis. Também aqui há factos morais óbvios – por exemplo: jamais se pode bombardear hospitais. Nenhum fim justifica tais meios. Em conflitos armados, políticos podem ordenar golpes "táticos" que violam factos morais, levantando a questão de uma ética dos *trade-offs*. Mas isso conduz-nos a um terreno difícil que, como defenderei por fim, só poderemos trabalhar com seriedade científica se usarmos sistemas de IA e outros métodos digitais.

## III. Inteligência Ética em vez de Ética da IA

Chego, assim, à terceira parte – mais breve do que as anteriores, porque terei ocasião de desenvolver estas ideias na minha Gulbenkian Lecture de 29 de outubro.

Uma tese importante que abordei hoje sob o título de "tese de Church-Turing" pode também ser resumida da seguinte forma: os limites da inteligência artificial são, neste momento, os limites da lógica formal clássica moderna. A lógica – de Boole e Frege – que subjaz à IA atual diz-nos que, em qualquer conjunto de dados digitalmente registável, os sistemas de IA conseguem identificar padrões melhor do que qualquer ser humano isolado. Quanto mais extensos forem os dados e quanto mais potente o *hardware*, mais a IA descobre padrões que a humanidade inteira, no seu conjunto, jamais encontraria. Segue-se que a IA é o mais poderoso instrumento epistémico que alguma vez desenvolvemos – e tornar-se-á, com avanços como a computação quântica e até *wetware* biológico, inimaginavelmente mais poderosa num futuro próximo.



Daqui resulta que os sistemas de IA são capazes de identificar melhor os nossos valores humanos nos conjuntos de dados que contêm vestígios do nosso pensamento e comportamento do que qualquer pessoa. Isso permite desenvolver sistemas que designo por *AlphaBuddha* – alguns já existem parcialmente. Eu próprio trabalho, numa *start-up* que fundei, em sistemas desse tipo.[12] No *Kyoto University Institute for the Future of Human Society*, por exemplo, agentes de IA inspirados sobretudo no budismo já se encontram em fase avançada, como pude observar há poucas semanas em Quioto.[13]

Aqui coloca-se a objeção habitual: dos dados que documentam pensamentos e comportamentos humanos não se pode extrair quais factos morais existem nem quais deveriam ser os nossos valores. Do ser dos dados não se infere um dever-ser moral.

Este argumento, porém, assenta num equívoco. Ignora que a investigação em economia comportamental, psicologia, neurociência e etnologia mostrou que os humanos se orientam sempre por questões morais e, portanto, deixam traços de juízos morais no seu comportamento efetivo. Uma das razões foi já apontada por Charles Darwin: somos mamíferos pró-sociais cuja coordenação de ações depende de nos comportarmos suficientemente bem, em termos morais, para sobreviver. Células-es-

[12] Cf. www.deep-in.ai
[13] Cf. https://ifohs.kyoto-u.ac.jp/en/

taminais humanas só se tornam, um dia, crianças capazes de sobreviver autonomamente porque adultos cuidaram delas – o que requer, no mínimo, cuidado afetivo. Neste contexto, Alva Noë, neurofilósofo em Berkeley, avançou recentemente que a consciência humana pode ser entendida como uma forma de amor. As formações sociais humanas – e, com elas, as sociedades – só funcionam porque consideramos, na esmagadora maioria das vezes, os factos morais adequados, sem necessidade de os fixar ao nível da reflexão ética. A ética manifesta-se naquilo a que Kant chamou o "modo de vida em concreto"[14].

O próximo passo da ética da IA é desenvolver agentes de IA capazes de extrair do nosso comportamento níveis novos de abstração que, a partir de factos morais óbvios, derivem respostas para questões morais complexas. Esses sistemas poderiam emitir recomendações falíveis, às quais responderíamos com reflexão ética e ação. Humano e máquina cooperam, neste modelo de Inteligência Ética (E.I.), para gerar um círculo virtuoso, uma espiral de progresso.



Pela primeira vez na era da IA, estamos em condições de o fazer. Por isso, a ética da IA deixará de consistir em pequenas comissões de filósofos e peritos a tentar, em gabinetes, regular o progresso tecnocientífico. Tal abordagem é pouco dinâmica para a IA e desperdiça as enormes oportunidades epistémicas da revolução da IA, que não se limitam, de modo algum, às ciências naturais ou às *digital humanities*. Surge, assim, um novo tipo de humanidades, que estamos a desenvolver no já mencionado *Institute for Philosophy and the New Humanities*, com foco na IA. Que conceito das ciências humanas, com as suas metodologias e formas de objetividade, e que tipo de cooperação transdisciplinar necessitamos para isso – é o tema que abordarei aqui dentro de algumas semanas, como base para precisar, de forma cientificamente rigorosa, o conceito de Inteligência Ética.

Limito-me a adiantar que a fórmula pessoana – "sentir tudo de todas as maneiras" – nos poderá guiar, representando, porém, apenas a superfície de um vasto património poético, poetológico e hermenêutico da esplêndida cultura lusófona. Estou ansioso por continuar a colaborar convosco aqui no Porto e espero que esta perspetiva de investigação seja inspiradora para todos – e para o ano letivo que agora se inicia. Por hoje, agradeço a honrosa oportunidade de proferir a Lição de Sapiência deste ano letivo.

---

[14] Immanuel Kant, *Fundamentação da Metafísica dos Costumes* (GMS, AA 4: 389).

ISBN 978-989-9193-75-8