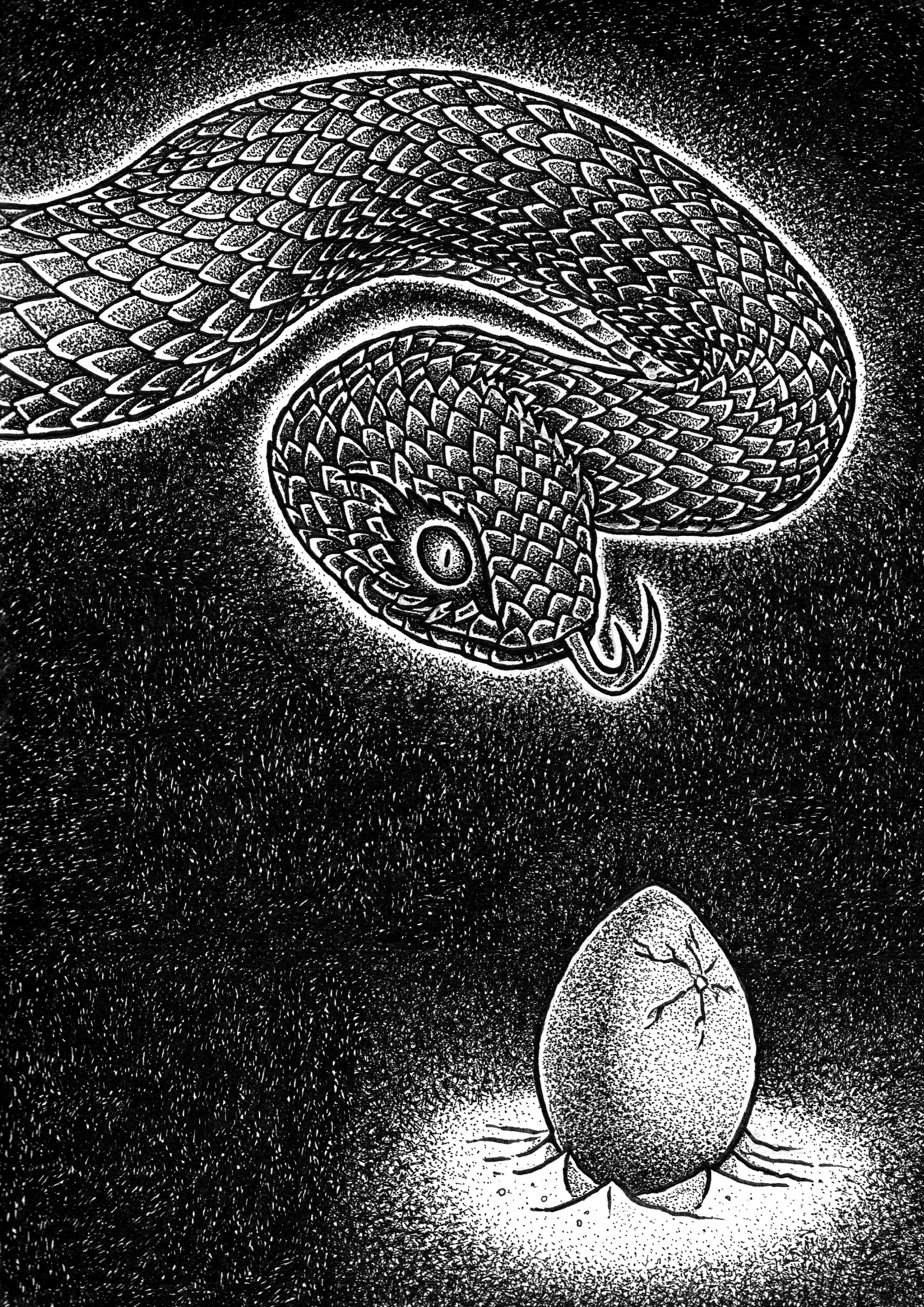

ACERCA DE MENTES NATURAIS E DIGITAIS, OU DE PROMESSAS ARRISCADAS

Luís Moniz Pereira

Departamento de Informática, Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa
2829-516 Monte da Caparica, Portugal
lmp@fct.unl.pt
ORCID 0000-0001-7880-4322

António Barata Lopes

ANQEP – "Agência Nacional para a Qualificação e Ensino Profissional" e
Agrupamento de Escolas de Alvalade
Av. 24 de Julho 138, 1200-771 Lisboa, Portugal
lopesab@msn.com

Resumo:

A Ciência da Computação diz-nos como criar um meta-interpretador para uma linguagem, escrito na própria linguagem, caso do Lisp e Prolog. A metalinguagem sendo igual à linguagem, faz com que ela goze da capacidade reflexiva de falar sobre seus enunciados e procedimentos. Da mesma forma, a Máquina de Turing básica permite fazer emergir, por *bootstrap*, a Máquina de Turing Universal. E portanto, permitir modelar qualquer computação máquinal. Como o Uroboros, a cobra que come a sua própria cauda, uma Máquina de Turing pode emergir para outros níveis de existência. Tais capacidades computacionais são boas demais para serem ignoradas pela

Abstract:

Computer Science tells us how to devise a meta-interpreter for a language, written in the language itself, viz. Lisp, Prolog. The metalanguage being the same as the language, makes it enjoy the reflexive ability to talk about its statements and procedures. Likewise, the basic Turing Machine lets emerge, by bootstrapping, the Universal Turing Machine. And thence permits to model any machine computation. Like the Uroboros, the snake that eats its tail, a Turing Machine can emerge onto other levels of existence. Such computational abilities were too good to be ignored by evolution, therewith fostering the emergence of cognition. Our brains, consciously and adeptly use and can teach

evolução, tendo promovido pois o surgimento da cognição. Os nossos cérebros, consciente e habilmente, usam e podem ensinar tais capacidades: ou seja, autodepuração, explicação e justificação, antecipação e preferência por futuros, visão contra-factual do passado com conhecimento do presente, actualização moral, detecção e remoção de contradições, argumentação, etc. A nossa tese: a computação, complementada por dados da Psicologia Evolucionária e por estudos sobre o comportamento de animais sencientes, é a ferramenta epistémica objectiva, evolutiva e abstracta, por excelência, para modelar a consciência, incluindo a evolução da moralidade nas populações em geral, empregando a Teoria dos Jogos Evolutivos. Esta abordagem epistémica computacional em direcção a uma teoria especulativa objectiva da consciência parece ser o actual paradigma *par excellence*.

Palavras-chave:

Inteligência artificial (IA); Alan Turing; máquina de Turing; Psicologia Evolucionária.

such abilities: namely self-debugging, explanation and justification, anticipating and preferring futures, counterfactually seeing the past with knowledge of the present, moral updating, contradiction detection and removal, argumentation, etc. Our contention: computation, complemented with data from Evolutionary Psychology and studies of sentient animals, is the objective, evolutionary, and abstract epistemic tool par excellence to model consciousness, including the evolution of morality in populations in general, employing Evolutionary Game Theory. The computational epistemic approach towards an objective speculative theory of consciousness appears to be the present paradigm *par excellence*.

Keywords:

Artificial intelligence (AI); Alan Turing; Turing machine; Evolutionary Psychology.

What's past is [*epistemic*] prologue, what to come in yours and my discharge.
William Shakespeare, *The Tempest* (1610-1611)
(Where Prospero bids farewell to black magic and enters a "brave new world")

Introdução

Quando pretendemos abordar conteúdos associados ao termo *mente*, dado que estamos perante um dos conceitos, ou ordem de conceitos, mais difícil de definir, temos sempre de considerar que enfrentamos um campo de investigação recente, e com muita margem para aprofundamento. Podemos, ainda assim, delimitar o seu âmbito a um conjunto agregado de processos de matriz cognitiva, emocional, e conativa, de natureza consciente e inconsciente, orientados para a aquisição, conservação, exploração, e avaliação de informações externas e internas.

Nestes termos, dispor ou não de uma mente não é, de todo, irrelevante. Embora os agentes destituídos dela possam estabelecer relações bem-sucedidas com o meio, o facto é que a capacidade de criar representações, internas e externas, é a condição de possibilidade para interacções mais diversificadas e complexas com o meio externo, e connosco próprios.

A ideia de estarmos perante um conjunto agregado de funções múltiplas não é nova no pensamento ocidental, podendo já ser encontrada no tratado de Aristóteles sobre a Alma[1]. Note-se que a ordem de conceitos com os quais caracterizamos a mente ancora no que tradicionalmente se designou por alma, espírito, ou outros termos próximos de uma concepção antropológica dual, sintetizando imanência e transcendência. Aquele filósofo grego tinha a noção clara de que tal capacidade não era exclusiva dos seres humanos, conjecturando, por isso, a existência de almas qualitativamente diferenciadas, e referindo-se também a uma gradação que, de alguma forma, quando actualizada para o presente, tem condições para ser lida como evolutiva; senão no tempo, pelo menos nas qualidades. É assim que Aristóteles nos fala de almas com faculdade sensitiva; sensitiva e locomotiva; e finalmente, sensitiva, locomotiva e intelectiva.

A temática das faculdades superiores da cognição foi sempre inerente ao discurso filosófico; mas, para construirmos uma noção mais dinâmica dela, foi preciso chegarmos

1 ARISTÓTELES. *Acerca del alma*. Introdução, tradução do grego e notas de Tomás Calvo Martinez. Madrid: Gredos, 1994.

ao século XIX, e a cientistas como Paul Broca, com o qual as faculdades mentais superiores – como o pensamento e a linguagem – começaram a ser vistas, não como aptidões decorrentes de uma misteriosa centelha divina, mas sim como produção de um órgão director que é o cérebro. Um cérebro resultante de pressões evolucionárias desafiantes, e de interacções com meios diversos, a exigirem respostas diferenciadas e complexas.

Ora, foi essa necessidade de diferenciação e sofisticação que veio a resultar na emergência da estrutura orgânica mais complexa, flexível e poderosa que a natureza engendrou. O cérebro encontra-se no topo de um corpo erecto, protegido por uma camada óssea extremamente robusta e rodeada de sensores de vária ordem, como se guardas de uma caixa-forte se tratassem. Por via dos membros superiores e inferiores do corpo onde está alojado, dispõe também de actuadores que, não só permitem interagir com o mundo, como "afinar-se" a si próprio. O seu comparativamente elevado custo energético assinala que sua evolução mereceu a pena.

1. Novas emergências paradigmáticas

Antes dos primeiros estudos científicos sobre o suporte biológico da cognição, tomando como base de investigação lesões ocorridas nos indivíduos e respectivas consequências, como os trabalhos conduzidos por Broca; ou partindo do estudo do comportamento animal, como os estudos conduzidos por Pavlov acerca da possibilidade de se condicionarem respostas comportamentais; ou ainda os primeiros passos da Psicologia Experimental conduzidos por Wundt, a anterior abordagem à mente humana tinha como pressuposto a existência em todos nós de uma dimensão espiritual, a qual era tida como incompatível com a sua abordagem por via das ciências experimentais, ou a construção engenheirística de artefactos sucedâneos. Postulava-se também uma ruptura entre humanos e natureza, segundo a qual os primeiros eram vistos como beneficiários da restante criação divina. A revisão dessa crença implicou que deixámos de nos percepcionar como a pedra angular da criação, para concebermos a vida e cognição de um modo integrado. Nessa integração, a mente humana ocupa, ainda e por enquanto, o topo da hierarquia mental; mas agora, enquanto resultado de um processo em evolução, o qual, atingido este patamar, permitiu que vida e a própria cognição tomassem consciência de si, interpelando não apenas a sua origem e natureza, mas também as suas potencialidades e limites.

Mudado o paradigma, presentemente, os avanços no domínio da Psicologia Evolucionária, e o conhecimento que vamos adquirindo sobre a cognição e comportamento

animal, permitem-nos construir uma visão muito mais coesa e integrada da evolução das espécies e da inteligência/cognição, enquanto processo igualmente evolucionário e distribuído. Além disso, toda a tecnologia aplicada ao universo das neurociências e das ciências cognitivas, como a imagiologia por ressonância magnética, ou engenharia de interfaces cérebros/computadores, entre muitos outros recursos, servem de base a conjecturas mais rigorosas acerca dos suportes biológicos para a inteligência e cognição.

Numa outra vertente, os desenvolvimentos da Inteligência Artificial (IA), associados à capacidade do que poderíamos designar por engenharia filosófica, permitem-nos simular e experimentar no computador processos que anteriormente apenas podiam ser objecto de especulação mediante reflexão e argumentação; embora, por vezes, com base em observações de campo, ou experiências intencionalmente conduzidas. Tais caminhos, devido ao seu potencial de explicitação e análise dos processos mentais humanos e não-humanos, permitem um olhar simultaneamente mais global e detalhado de todos os processos cognitivos. Mais global, porque nos remetem para a mente enquanto faculdade detida por várias espécies, quiçá extraterrestres; e também mais detalhado, pois dispomos de conceitos e instrumentos de investigação que nos permitem "entrar" no domínio da mente em acção, por assim dizer. Mesmo a moral, vista como caso de investigação com recurso à teoria dos jogos evolucionários, é actualmente, objecto de abordagem por via da Psicologia Evolucionária e das Ciências da Computação.

Um dos aspectos que primeiramente devemos abordar diz justamente respeito ao carácter evolucionário dos processos mentais, o qual deve ser considerado em várias vertentes. Desde logo, explorando o facto de partilharmos tais faculdades não apenas com os outros símios, mas também com golfinhos, elefantes e todos aqueles que detêm um sistema nervoso suficientemente complexo para construir representações do meio e linguagens que as expressem (incluso os eventuais extraterrestres); ressalvando, evidentemente, que essas linguagens de outras espécies (pelo menos as terrestres) não têm o potencial das humanas, as quais representam objectos na sua ausência, instauram mundos possíveis, imaginam mundos contra-factuais, diversificam a simbologia, e desdobram-se em meta-linguagens.

É claro que, presentemente, estamos muito longe de Aristóteles e, por exemplo, da consideração segundo a qual apenas os humanos, meta-linguisticamente considerados, riem. Todavia, o riso, também reconhecido noutros símios, é tão só um mecanismo de ligação entre seres com a faculdade de sentir e transmitir afectos. Assim, uma mãe chimpanzé, ou uma mãe humana, ao fazerem cócegas aos seus bebés, obtê-lo-ão como

resposta. Podemos, pois, estabelecer uma conjectura razoável segundo a qual o riso humorístico humano tem uma base ancestral e distribuída por mais outras espécies, para além dos próprios. De igual forma, outras faculdades humanas mais complexas e que exigem uma mente com a capacidade de construir representações abstractas do mundo são também o resultado de competências emergentes, mesmo que tivessem o seu início de modo muito mais rudimentar.

Dois casos estudados por Franz de Waal[2] permitem-nos ilustrar com clareza o que acabámos de afirmar. O primeiro diz respeito às investigações conduzidas com grupos de elefantes asiáticos, na Tailândia, as quais permitem não apenas evidenciar o forte sentido comunitário desta espécie, bem como comportamentos associados a um rudimento de moral, quer em situações de entreajuda entre pares, quer também no modo como os adultos engendram estratégias de protecção das crias face a ameaças.

Foram também investigadas todas as estratégias de comunicação vocal e não vocal (linguagem corporal) que permitem consolidar o sentido de comunidade, a linguagem complexa, as capacidades de avaliar situações de risco individual e colectivo, a construção de estratégias concertadas e as trocas de expressões emocionais após a vivência e superação de situações de risco. Os elefantes em cativeiro, usados como força de trabalho, manifestam inequivocamente comportamentos projectivos, como introduzir forragem num chocalho, a fim de se poderem deslocar silenciosamente, em passeios nocturnos, à revelia dos seus guardadores humanos; e têm os antepassados em consideração por via de um "culto" da morte, que os leva a visitar o local onde pereceram os membros da sua família, como forma de solidificar o intra- e o inter- geracional.

O outro caso diz respeito aos estudos conduzidos por Sarah Brosnan e Frans de Wall, no Brasil, onde os indivíduos testados foram macacos-prego[3]. Essas experiências permitiram validar a capacidade destes símios para avaliação de certas tarefas, e comparação das respectivas recompensas. Foi também possível indagar se o seu comportamento seguia critérios que, minimamente, se pudessem equiparar a um rudimento de moral.

A situação experimental, protagonizada por fêmeas da espécie, consistiu em colocá-las perante a troca de uma simples pedra por uma rodela de pepino. A partir do momento em que um dos sujeitos experimentais começa a ser recompensado com bagos

2 Plotnik JM, de Waal FBM, Reiss D. 2006. Self-recognition in an Asian elephant. *Proceedings of the National Academy of Sciences of the United States of America* 103:17053-17057

3 https://www.youtube.com/watch?v=JJsKS4DQ95E

de uva, ao passo que o outro continua a receber rodelas de pepino, emerge a percepção da desigualdade. Então, após algumas repetições do mesmo procedimento, o sujeito experimental recompensado com rodelas de pepino opta por recusar o alimento, chegando mesmo ao ponto de o arremessar aos cientistas.

Tal comportamento prova que estes símios são capazes de monitorizar a sua acção e a dos parceiros, de analisar situações relativas, de as percepcionar como injustas porque desiguais, e de construir uma resposta coerente com a sinalização negativa em questão. Não nos devemos esquecer que valorizar é sempre estabelecer uma relação entre itens, a qual é independente dos conteúdos específicos de cada um deles; ou seja, implica sempre um constructo mental, qualquer que ele seja, elaborado pelo sujeito avaliador, no âmbito do qual valoriza uns aspectos em detrimento de outros.

2. A evolução distribuída emergente

Os riscos associados à projecção de características humanas em animais estão suficientemente documentados na literatura científica; no entanto, o pretexto de evitarmos tais riscos não deve, por outro lado, coibir a análise. Ora, as situações experimentais atrás mencionadas evidenciam que a mente é um fenómeno evolucionário, inter-espécies, regida por uma dinâmica de complexificação progressiva. Nos *Homo sapiens*, essa mesma dinâmica permitiu criar representações dos sujeitos em acção, de matriz colectiva e individual, com respectiva avaliação e revisão de estratégias. Essa complexidade está apoiada em linguagens com níveis de abstracção – e inerente artificialidade – cada vez mais diversa.

Desta forma, o jogo evolucionário ocorre quer ao nível dos organismos vivos que vão emergindo, quer ao nível das faculdades que lhes são próprias, incluindo a capacidade de modificar e adaptar-se ao seu eco-nicho. Cabe aqui a introdução do termo "artificialidade", pois, num certo sentido, os suportes dos vários tipos e modos de inteligência podem ser vistos como o hardware, que serve o exercício das funções já referidas; por outro lado, os conteúdos mentais e respectiva organização, que são constituintes de cada mente, podem ser vistos como software.

Nesta linha de pensamento, até dada a reconhecida fragilidade dos nossos suportes biológicos, podemos muito bem conjecturar que um dos maiores desafios enfrentados pelos humanos será proceder à migração de funções cognitivas e, por inerência, mentais, para suportes mais robustos e duráveis relativamente a um corpo biológico sujeito às

agressões de vírus e bactérias, e – por enquanto – com poucos recursos para enfrentar a inevitabilidade da degradação e da morte.

É nesta base que podemos conjecturar a existência de futuras mentes digitais, ou bio-digitais, mercê da hibridização entre organismos biológicos e artefactos de IA. As próteses para sensores, actuadores, e outros órgãos do nosso corpo serão, certamente, complementadas com próteses cognitivas que originarão humanos melhorados, e mesmo simbiontes distribuídos. Não sabemos quais as consequências psicológicas, sociais e económicas de tal caminho. Mas sabemos muito bem que os humanos nunca deixaram de experimentar tudo o que é tecnicamente possível. Acresce que os desafios que colocamos a nós próprios, como sejam a exploração espacial, a manipulação genética, ou a gestão de sistemas e plataformas globais progressivamente mais sofisticadas, multifuncionais e integradas, têm o seu actual desempenho e progresso imputado ao desenvolvimento da IA.

As questões associadas à gestão e integração de dados, com a migração do conceito de aprendizagem para o domínio das Ciências da Computação, autorizam que, por inerência, se tenha iniciado e se desenvolva uma nova literatura em torno da ideia de mente digital. Não que estejamos próximos do conhecimento artificial auto-consciente, mas apenas porque já o podemos conjecturar, e não o eliminar *a prori*. Os humanos sempre especularam sobre a sua relação com inteligências outras; referimo-nos a deuses e a anjos, mas também podemos invocar extraterrestres ou monstros que, sorrateiramente, pudessem coabitar connosco, no mesmo planeta. Mesmo a possibilidade de criarmos autómatos nos quais fosse possível injectar vida esteve sempre presente na especulação humana[4]. Assim sendo, é, e não é, surpreendente que – com quase toda a certeza – a primeira inteligência-outra com a qual interagiremos será uma criação nossa, mas suficientemente empoderada para ser autónoma.

Nestes termos, antecipamos a emergência de um ecossistema cognitivo onde seres biológicos não humanos, humanos melhorados, eventualmente extraterrestres, e máquinas cognitivas formarão um *cluster* destinado à produção e partilha de conhecimento altamente diversificado e distribuído. Sendo que essa diversidade propiciará, certamente, melhores porque mais diversificadas representações tanto do mundo como das próprias modalidades de mente. Sendo também possível imaginarmos a construção de uma mente que aglutine funcionalidades oriundas de domínios diversos.

4 "*AI Narratives – A History of Imaginative Thinking about Intelligent Machines*". Edited by Stephen Cave, Kanta Dihal, Sarah Dillon, Oxford University Press, Oxford, UK 2020.

Presentemente, damos pequenos passos que permitem grandes avanços em várias direcções, nomeadamente no domínio da modelização matemática vectorial que suporta a Linguística Computacional e que, a breve trecho, interferirá drasticamente em todos os domínios das relações humanas. Por outro lado, a alvorada da computação quântica promete-nos saltos qualitativos e quantitativos no tratamento de dados que, por enquanto, apenas conseguimos vislumbrar.

3. Construir para compreender

Richard Feynman tem uma frase famosa: "A prova da compreensão está na construção". Supomos que uma abordagem complementar em direcção a um modelo abstracto da consciência reside na resolução computacional de problemas comuns reconhecidos por exigirem consciência, incluindo processamento reactivo e deliberativo. Partindo então das inovações computacionais que tais problemas e suas combinações nos levem a alcançar, um de nós empregou essa abordagem para modelar a moralidade, tanto para o raciocínio moral individual (começando com os problemas do bonde) quanto para modelar o surgimento e a evolução da moralidade em populações de tais indivíduos, incluindo o sentimento de culpa (Pereira *et al*. 2017; 2023)[5].

Somos a favor de uma postura funcional de Turing em relação à consciência, ou seja, prontamente vista como implementável de acordo com Feynman. No entanto, absorvendo inspiração experimental humana/animal/insecto, consciência distribuída incluída. Com respeito à cognição, enfatizamos todos os 3 níveis indispensáveis detalhados em Pearl (2018)[6], em consonância com o Tri-Processo (Stanovich, 2010)[7]: previsão e reacção por correspondência e aprendizagem; hipótese de cenários com abdução e planeamento; contra-factualizando o passado com o conhecimento presente.

5 T. Cimpeanu, L. M. Pereira, T. A. Han. Co-evolution of social and non-social guilt in structured populations, extended abstract, in 22nd Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (Eds.), *Proceedings ACM Digital Library*, London, UK. May 29 - June 2, 2023.
L. M. Pereira, T. Lenaerts, L. A. Martinez-Vaquero, T. A. Han. Social Manifestation of Guilt Leads to Stable Cooperation in Multi-Agent Systems, in: *Procs. 16th Intl. Conf. on Autonomous Agents and Multiagent Systems* (AAMAS 2017), Das, S. et al. (Eds.), pp. 1422-1430, 8–12 May 2017, São Paulo, Brazil.

6 Judea Pearl, Dana Mackenzie. "*The Book of Why: The New Science of Cause and Effect*". Basic Books, New York, NY, USA 2018.

7 Keith Stanovitch. "*Rationality and the Reflective Mind*". Oxford University Press, Oxford, UK 2010.

No entanto, mesmo considerando todos os actuais avanços, estamos ainda muito longe de criar uma mente digital; embora essa limitação, que diríamos circunstancial, não nos impeça de implementar na máquina funcionalidades que, no passado, eram exclusivas da mente humana. Sucede *apenas* que as máquinas não têm consciência das tarefas que estão a realizar, nem as articulam numa sinfonia cognitiva bem composta. Podemos, por exemplo, programar a geração de contra-factuais; criar simulações de natureza estatística que nos permitam validar os efeitos de comportamentos egoístas, ou altruístas, ou de um misto de altruísmo e egoísmo devidamente doseado; ou ainda avaliar os efeitos do acto de pedir desculpa no contexto da moral dos grupos.

Tais aspectos são ganhos de investigação alcançáveis no contexto do conhecimento sem consciência, propiciado pela actual IA. Devemos reconhecer que estamos limitados não apenas pelo actual poder de computação, mas fundamentalmente porque ainda não conhecemos muito bem os processos que fazem a consciência emergir em certos organismos vivos; também não conhecemos suficientemente os recursos e procedimentos através dos quais a nossa mente constrói mapas internos e externos, por forma a interpretar o presente e antecipar o futuro, e temos um conhecimento fragmentado dos nossos mecanismos morais.

A mente humana, apoiada num entrelaçado de raciocínios e emoções, tem esse carácter antecipador, não apenas porque gostamos de construir melhores explicações, mas fundamentalmente porque sabemos que as nossas possibilidades de sobrevivência estão, como sempre, muito dependentes dessa capacidade de antecipar futuros possíveis; e, muito provavelmente, se tivermos tal poder, de influenciarmos no sentido do que mais nos convém e sabermos evitar e combater os seus maus usos.

4. Perigos e precauções

Presentemente, há um *chatbot* de conversação que sinaliza de vez a presença cada vez mais determinante da IA no nosso quotidiano; referimo-nos ao ChatGPT4, uma criação da empresa Open AI. Este Chat GPT tem a particularidade de interagir connosco numa dinâmica de perguntas ou comentários e respostas; não apenas alimentado por dados existentes na internet, como também pelas modalidades de interacção que estabelece com humanos. Num diálogo aprofundado, o Chat GPT não passa o teste de Turing e outros testes que se lhe põem no caminho; mas, quando se trata de fornecer definições de conceitos, teorias, comparações entre teorias, e muitas outras acções baseadas em conhecimento já padronizado e bastante difundido, apresenta um desempenho assinalável.

No entanto, devemos ter cuidado com os seus limites, e os dos chats seus similares, pois enquanto nos ajuda com respostas que vão sendo reforçadas pelos utilizadores, também reduz o leque de possibilidades de abordagem a um problema ao que cabe dentro dos padrões maioritariamente validados. No limite, as ferramentas que trabalham grandes massas de dados a fim de encontrar padrões acabam por reduzir as alternativas aos padrões dominantes. Gostamos da facilidade, esquecendo com demasiada ligeireza que estamos perante máquinas de colectar dados; no caso do Chat GPT, reproduzindo as expressões linguísticas mais comumente associadas na miríade de documentos *consultados*. Ainda assim, dado que na maior parte de nosso tempo lidamos com conhecimento padronizado, é possível que a máquina nos crie a ilusão de ser um interlocutor genericamente credível, nomeadamente pelo seu bom domínio gramatical, comparável ao humano.

Para aprofundarmos um pouco mais os efeitos múltiplos da tecnologia em questão, e de outras que brevemente surgirão no mercado, podemos tomar como termo de comparação a disseminação das máquinas de calcular. O efeito da massificação do uso de tal instrumento foi a externalização de uma capacidade que, anteriormente, apenas podia ser desempenhada por um cérebro humano. Concretizado este processo – actualmente, parte do menu de qualquer smartphone de gama baixa – os humanos passaram a poder fazer cálculos complexos com muito maior rapidez e segurança; mas a consequência, porventura não desejada, foi a diminuição drástica do adestramento dos cérebros humanos nesse domínio. É certo que aprendemos a calcular na escola, mas também não há dúvidas de que, sem exercícios futuros, acabamos por esquecer certas regras segundo as quais se executam essas operações. O resultado, no que concerne à faculdade específica de calcular, é a nossa dependência das máquinas.

Do ponto de vista estritamente humano, não diremos, ainda assim, que perdemos muito; pois estamos a considerar uma competência mecânica, que pouco acrescenta à flexibilidade multifuncional que se espera de um cérebro humano. Acresce que os riscos decorrentes da perda de tal faculdade são baixos, já que perto de nós existirá sempre uma máquina de calcular! Por contraposição, um chatbot como o GPT, ou outro sucedâneo mais desenvolvido, tem capacidade para substituir competências – a exemplo da do cálculo – mas num vasto espectro de funções permitidas pela nossa mente, e expressas em linguagem natural. Referimo-nos a sínteses, explanações de teorias, comparações, e muitas outras capacidades que apenas realizamos por termos uma mente.

Ora, mesmo sem ser de facto uma mente, uma ferramenta como o ChatGPT, operando sobre dados já produzidos, pode actuar como tal e apresentar resultados muito

aceitáveis. Combinando os dados existentes, pode ajudar a criar compostos químicos inovadores que sejam melhores medicamentos, novas terapêuticas, ou evidenciar padrões comportamentais que antes nos escapavam. Em suma, a capacidade de tratar e retirar informação de grandes massas de dados e as combinatórias exploradas através de estratégias diversas—a exemplo do que nós, os humanos fazemos, mas sem a capacidade de lidar com o mesmo volume—são um prometedor campo de aplicação das máquinas que usam os modelos LLM (*Large Language Models*).

Neste sentido, estas tecnologias têm capacidade para, ou gerar dependência, ou para se tornarem parceiras, em áreas muito mais significativas para os humanos do que o cálculo. Ou seja, usadas sem critério e massificadas, comprometerão a nossa disposição para construir conhecimento e desenvolver pensamento crítico, detecção de contradições e procura de fontes, por, aparentemente, nos substituírem nessas funções, já que os actuais chatbots não têm capacidade crítica de opinião nem sobre a credibilidade das suas fontes (que até podem, recursivamente, serem outros chatbots). Por outro lado, devidamente contextualizadas e integradas a título de ferramentas, serão um excelente auxiliar para agregação de informação sobre a qual a mente poderá exercer as suas superiores faculdades.

É neste contexto que realçamos o facto de parte substancial das nossas aspirações sobre o conhecimento estarem relacionadas com a necessidade de antecipar e moldar o futuro. Ora, nesse domínio uma máquina que agrega dados e encontra padrões neles; nomeadamente, identifica as palavras que, estatisticamente consideradas, melhor se articulam com as anteriores, seleccionando a mais adequada, será de pouca utilidade. Pensar fora dos padrões, ou – de modo mais ousado – pretendermos usar o conhecimento que detemos como base para a criação de novos paradigmas, exige o uso de uma mente livre e ousada; capaz de colocar e explorar hipóteses que, dadas as suas características emergentes, não encontraremos nos padrões já estabelecidos.

Uma sociedade maquinal, altamente padronizada, normalizadora dos comportamentos, apreciará infelizmente as competências de tais máquinas e de mentes humanas moldadas por elas, restringindo-lhes por habituação estatística o vocabulário e conceitos nele expressos, à maneira da língua padronizada *Newspeak* no livro "1984" de George Orwell. Mas, nesse território, temos de ter a noção de que estamos já a contextualizar o que será uma ciberditadura, usando a máquina a título de delimitação dos âmbitos, e instrumento de controlo. Como se tivéssemos chegado ao fim da História, tendo agora como desígnio gerir e optimizar os recursos cognitivos adquiridos. Cumulativamente, surge um

novo neo-colonialismo cultural, uma vez que estas tecnologias tratam em número desigual línguas e alfabetos de todo o mundo, por desigualdade na origem dos documentos rastreiam.

Reproduzidos via ensino, os fenómenos perversos chegarão rapidamente às novas gerações, as quais terão menos sentido crítico e menos aptidão para verificar as fontes de informação. O recurso à informação estatística para formulação das suas próprias opiniões nas redes sociais converterá estes espaços em diálogos entre "GPTs" configurados de forma pessoal, como avatares que nos vão substituir. E estes são ingredientes perfeitos para potenciar o crescente afunilamento conceptual, de vocabulário e de opiniões. O cruzamento de fontes será o caminho para restaurar o pensamento crítico. Caso contrário, os humanos terão mentes cada vez mais previsíveis e formatadas nas respostas que vão dar.

O lançamento do ChatGPT foi como se disponibilizasse um novo avião que não foi avaliado e certificado. Isto acontece porque a legislação actual assume que o "software" não precisa disso. Mas precisa! No actual ambiente legislativo corre-se o risco de se lançar um vírus cognitivo com consequências importantes e não avaliadas previamente. A disponibilização destes instrumentos é como um "tsunami" que, como em quase tudo na informática, se propaga de forma imediata, por todos e por toda a parte. Acresce que os investigadores não estão habituados a este nível de avaliação de grandes impactos sociais da tecnologia. Pelo contrário, entendem que há uma oportunidade para arrecadarem mais recursos, abstendo-se de opiniões negativas globais, perpetuando dessa forma a prática de se esconderem os eventuais problemas debaixo da carpete. Com este contexto, a pergunta que subsiste é: quem beneficia lucrativamente com a IA? Essa questão não é trazida para a superfície. A IA deveria originar uma maior riqueza adquirida para todos, e não concentrar o lucro apenas em alguns.

A propósito, não podemos deixar de fazer uma breve alusão ao impacto destas novas tecnologias no mercado de trabalho. Dado que as profissões humanas, na sua maioria, são rotineiras, exigindo poucas capacidades ligadas à componente criativa e crítica da mente, máquinas rotineiras estarão em condições de substituir muitos humanos, sem que exista por agora uma reflexão suficientemente articulada sobre a matéria. Resumidamente, quer do ponto de vista colectivo, quer individual, devemos ter a noção de que as tecnologias da IA são tanto uma oportunidade, se devidamente enquadradas, como um risco, se abdicarmos das nossas faculdades e as substituirmos pelos seus aparatosos, mas ao mesmo tempo fracos recursos. Reiteramos que mentes digitais são, por agora, uma miragem e que os seus simulacros não agem por conta própria, mas ao serviço de reais men-

tes cujo interesse passa por funcionalizar os seus semelhantes, desvalorizar o trabalho, e concentrar a riqueza que será cada vez mais produzida por máquinas.

A IA está já a gerar impactos sociais graves em níveis cognitivos inferiores, mas muito rapidamente evoluirá para novos patamares cognitivos. Uma IA simples pode ser mais perigosa do que uma IA complexa, por ser meramente estatística. Numa interrogação ao ChatGPT, ele vai verificar as sequências de palavras/resposta mais prováveis na documentação que se relaciona com as palavras da interrogação. Ou seja, é um sistema que mede as coisas em termos estatísticos e que acaba por assumir que o futuro será igual ao que já está fixado do passado. Este nível, o do reconhecimento de padrões, é considerado o mais baixo da IA. Esta simplificação tem um lado perigoso e ao mesmo tempo perverso. Porque o ser humano, ao tornar-se muito dependente dela, treina cada vez menos o pensamento crítico, e a capacidade de avaliar a credibilidade e consistência da informação fica diminuída. Num nível superior de IA teríamos as hipóteses e os cenários possíveis, bem como o patamar confra-factual. A característica deste é o humano perguntar-se: «Que aconteceria se eu no passado tivesse feito de outra maneira?»

5. Da moral do homem à moral da máquina

Mas o que existe é o actual contexto, e nele interessa sobremaneira continuarmos a aprofundar o conhecimento sobre a mente humana, os seus processos e potencialidades. Esse conhecimento é a base para podermos implementar funcionalidades específicas em máquinas; esperando que futuramente seja possível uma melhor integração funcional. Por agora, como já mencionámos, um dos maiores riscos associados à IA é o de estarmos a delegar demasiadas funções, com os poderes a elas associados, em máquinas cuja inteligência é rudimentar.

De facto, a interacção com máquinas e a delegação de poderes nelas implica confiança; isto é, embora certos de que elas não têm uma mente crítica como a humana, temos de esperar que, ainda assim, tomem decisões de acordo com critérios aceitáveis para os seres humanos. Eis mais uma razão para suportar a tese que temos defendido, segundo a qual um dos domínios que urge investigar e aprofundar respeita a um melhor conhecimento da moral humana, os seus princípios fundamentais e o impacto deles nos grupos. É certo que estamos ainda numa fase muito embrionária, mas o esforço justifica-se porque não é possível dispormos de maquinaria cognitiva, com graus de autonomia cada vez maior, sem nos certificarmos que as mesmas são confiáveis. Acresce que o contexto onde

as máquinas têm de tomar decisões é, amiudamente, complexo; podemos tomar a guerra como exemplo, mas também os cenários decorrentes de catástrofes naturais, ou a necessidade de decidir no momento se se compram, ou vendem, certos lotes de acções na Bolsa; tal é o espectro de aplicação da IA.

Ora, esta investigação só é possível porque deixámos de percepcionar a moral como uma dádiva divina e, dentro do paradigma das investigações já referidas, olharmos para esse fenómeno evolucionário enquanto cola agregadora dos grupos e elemento fundamental da sua coesão. É assim que, presentemente, se investigam elementos presentes em todos os sistemas morais (independentemente da sua origem cultural), como é o caso do sentimento de culpa e do pedido de desculpa, ou do dilema cooperar/competir.

Quer em termos académicos, quer em termos empresariais, quer também em termos políticos, os avanços da IA colocaram a ética no centro da discussão. Mesmo na Meca do liberalismo, os EUA, os impactos possíveis de uma ferramenta como o ChatGPT levam a que empresários e outros cidadãos clamem pelo papel regulador do Estado. De entre os temas considerados como prioritários no âmbito da moral computacional, há a destacar a questão da responsabilidade algorítmica, onde se discute quem assume as consequências de acções e decisões tomadas por algoritmos e sistemas de inteligência artificial. Questões como o enviesamento algorítmico, a transparência na recolha de dados, a explicabilidade do processo de decisão, e a responsabilidade moral e legal por agentes artificiais são de natureza crítica; daí que a questão de como garantir que os sistemas de IA sejam projectados para respeitar a integridade humana, evitar fraudes, competição desregulada, e entrada no mercado de produtos sem os devidos testes de segurança sejam urgentes.

Em suma, referimo-nos sempre a questões que, anteriormente, estavam sob a exclusiva responsabilidade de mentes humanas. Este processo de delegação de competências é tão ancestral quanto a própria humanidade. O esforço físico árduo e repetitivo desumaniza-nos; por isso, os mais fortes escravizam os mais fracos e inventamos tecnologias que amenizam essas tarefas. Enquanto isso, trabalhamos os músculos no ginásio, mas essencialmente por razões estéticas, ou, na melhor das hipóteses, de saúde. Da mesma forma, o esforço intelectual é dispendioso em matéria de concentração e dispêndio de energia, de modo que aceitamos com tranquilidade, e até ansiedade, que as máquinas nos substituam nessas tarefas. E, nesse processo, não reflectimos suficientemente sobre os seus modos de concretização sem, simultaneamente, nos desumanizarmos.

Numa outra linha de problemáticas, encontram-se as questões associadas à recolha massiva de dados, e os impactos na privacidade e segurança dos utilizadores da internet e dos recursos digitais em geral. Se os dados são a matéria-prima mais valiosa do Século XXI, cumpre perguntar porque é que os seus produtores não recebem por eles. É certo que, não existindo mentes digitais capazes de, por elas próprias, transformar dados em conhecimento, não há o perigo de a nossa mente ser imediatamente controlada por máquinas. O que não significa que não o possa ser pelos seus donos ou pelos utentes dos serviços prestados pelas máquinas cognitivas.

Dada toda esta ordem de problemas, o desafio de se desenvolverem algoritmos que incorporem critérios morais nos seus processos de decisão, e que sejam capazes de os explicitar em justificações, é de fundamental relevância. Continuaremos a não ter, ainda assim, verdadeiras mentes digitais como parceiras, mas poderemos aspirar a mais confiança no apoio que é dado ao desenvolvimento da nossa. De facto, presentemente dispomos de algoritmos para desempenhar funções delimitadas, como pilotar um drone ou colectar dados sobre matérias específicas, encontrando aí padrões que escapavam ao olhar humano.

Dada esta especialização, que é, simultaneamente, uma severa limitação, especula-se sobre o que se designa por singularidade tecnológica, ou seja, esse momento hipotético no qual a máquina superará o humano concretizando o que se designa por Inteligência Artificial Geral (IAG). Para já, não estamos próximos desse momento, dado que, embora conheçamos com bastante detalhe cada um dos nossos processos cognitivos, ou o papel das emoções, estamos muito longe de nos pronunciarmos com detalhe sobre a sua integração num todo a um tempo consciente e inconsciente, e o modo como o nosso cérebro com o seu corpo, em interacção com os outros cérebros e corpos fazem isso. Aumentar a velocidade e a escala no que diz respeito a tratamento de dados não permitirá, em princípio, obter ganhos em termos qualitativos.

A forma mais complexa de inteligência que temos ao nosso alcance para estudo e aplicação em outros suportes é a humana, daí que os desafios estruturais da IA estejam muito dependentes do desenvolvimento das Ciências Cognitivas; estas, dada a sua abordagem interdisciplinar da mente e da cognição, interceptam a Filosofia, a Psicologia, a Inteligência Artificial, as Neurociências, a Linguística, e a Antropologia. O seu desenvolvimento, nas várias áreas que a compõem, é de crucial relevância para suportar a emergência de melhores modelos de cognição artificial, ancorados num melhor conhecimento das mentes biológicas e dos seus processos. De entre toda a panóplia de problemas que

constituem a sua agenda, gostaríamos de referenciar a emergência de um novo inconsciente colectivo, de natureza digital, moldável pelas preferências dos utilizadores e com um incomensurável potencial não só para condicionar os comportamentos humanos, mas também para influenciar o domínio emergente da Ética da Máquina. E é com esse tópico que gostaríamos de terminar o presente artigo.

6. Moralidade humana e ética da máquina

A ética da máquina questiona como projectar, implementar e tratar robôs; que capacidades morais devem eles ter, e como concretizar cada uma delas. Dada a complexidade do tema, em vez de almejarmos fixar todos os critérios para a competência moral de um robô, podemos ter como objetivo identificar certos elementos da moral humana e, em seguida, investigar o desenho da máquina considerando selectivamente alguns deles.

Para concretizar tais desideratos, temos de ter em conta que os robôs poderão interagir entre si, mas sempre ao serviço de algum interesse humano, grupal ou individual. Daí que os seus padrões de moralidade tenham de estar em simbiose connosco. Ora, a moralidade desenvolveu-se durante a evolução. Somos uma espécie gregária, o que implica ter regras de convivência. Por outro lado, não existe uma teoria universal da ética, mas uma combinação de teorias éticas: Categórica; Construtivista; Utilitária; de Virtude; etc. Daí o ser problemático que não conheçamos a nossa moral suficientemente bem, e com detalhe, para que ela possa ser prontamente programada. Dado esse constrangimento, devemos começar por programar as normas que já temos bem padronizadas e definidas em contextos específicos, como sejam o hospitalar, as bibliotecas, as casas de repouso, a negociação financeira, os parques de diversões, os centros de compras, ou os teatros de guerra. Nunca podemos perder de vista que estamos apenas no início da ética de programação para máquinas.

Ainda considerando essa fase embrionária, é urgente que se conheçam melhor as facetas morais humanas como sejam o vocabulário moral, as normas morais, a cognição moral e o afecto, a tomada de decisão moral e a sua relação com a acção, ou a comunicação moral e o consentimento. O seu estudo profundo é um pré-requisito para se progredir no ADN ético da máquina. Algumas capacidades que servem de substracto ao comportamento ético nem precisam de linguagem verbal; referimo-nos ao reconhecimento de comportamentos prototipicamente pró-sociais e antissociais, ou à empatia e reciprocidade básicas. Já para outras, torna-se necessário um vocabulário padronizado a fim de que

possamos aprender, ensinar e deliberar sobre elas de modo bem definido. Não devemos deixar fora do processo de explicitação e definição todo o vocabulário para expressar práticas morais como sejam as decisões de culpar ou desculpar um comportamento, perdoar, justificar, ou negociar as prioridades duma norma.

Além disso, necessitamos de um vocabulário para as próprias normas: nomeadamente no que respeita aos termos *justo, virtuoso, recíproco, honesto, obrigatório, proibido, desejável* etc. E ainda de explicitar o que entendemos por palavras que exprimem violações de normas como sejam *errado, culposo, imprudente, ladrão, intencional, conscientemente, acidental etc*.; bem como de resposta às violações de noemas, como sejam *culpa, repreensão, desculpa, perdão etc.*

Para formar juízos dirigidos a agentes, como seja a culpa, um robô precisa de capacidades para raciocínio causal sobre eventos segmentados, Inferências sócio-cognitivas do comportamento para determinar intencionalidade e razões, além de raciocínio contra-factual para decretar a prevenção. Note-se que uma componente proeminente da competência moral humana é a tomada de decisão e respectiva articulação com a acção. Neste contexto, a culpa é pedagógica pois é o modo de fornecer ao infractor da norma razões para não repetir. A culpa regulará o comportamento do robô se ele aprender a levá-la em consideração, nas suas próximas escolhas. Já o livre-arbítrio metafísico não é, de todo, necessário.

Na projecção de robôs capazes de decisões e ações morais, a tensão entre o interesse próprio e os benefícios da comunidade deve ser evitada desde o início. Deve ainda ter-se em conta que os robôs de diferentes fabricantes irão competir entre si! O tipo de robô que imaginarmos não pode ser programado para agir moralmente em todos os futuros possíveis. Terá normas orientadoras no início, mas precisa aprender a reformulá-las e a seleccionar outras. Portanto, corre o risco de deixar de agir moralmente por ignorância. Daí que o feedback seja fundamental para ele poder fazer melhor na próxima vez. No entanto, algumas situações apresentam problemas de decisão em que nem todas as normas relevantes podem ser satisfeitas em conjunto. Tais dilemas morais exigem uma escolha genuína entre opções imperfeitas.

Noutras ocasiões, cada opção pode ser moralmente justificada por referência a normas aceitáveis. Daí que as ferramentas cognitivas para juízo moral e tomada de decisão, por si só, sejam insuficientes para a função social de regular o comportamento dos outros. Ou seja, por vezes disposições humanas como a compaixão, ou a empatia, são critérios utilizáveis em ordem à boa decisão moral. Sendo assim, os robôs precisarão ganhar

flexibilidade de capacidade de reconhecimento de disposições humanas, geradores de um nível de confiança tal que lhes permita monitorizar e fazer cumprir as normas. Devem ainda declarar a obrigação de denunciar infrações às normas, e utilizar a comunicação para alertar e lembrar os preceitos aplicáveis.

As simulações envolvendo IA são ainda um veículo privilegiado para ensinar e treinar comportamentos morais, de forma interactiva com humanos, através de jogos morais. Tais jogos de computador podem ser empregues para testar teorias éticas e melhorar a educação moral, por meio de exemplos e explicações. Esses jogos podem contribuir com ferramentas para conceber, gerar e ilustrar comportamentos morais interactivos, em jogos multi-jogador, individuais e colectivos (Pereira 2022)[8].

E se, no processo de diversificação da oferta, surgirem máquinas com moral incompatível? Este é um cenário plausível, pois diferentes fabricantes produzirão máquinas com software moral distinto. Esse cenário pode vir a revelar-se caótico, pois as máquinas precisam de cooperar entre si por meio de uma moralidade comum, em vez de competir fora da ética, havendo o risco de robôs deliberadamente programados com intenções sinistras. Um objetivo importante da moralidade é, pois, a detecção de intenções desfavoráveis, trapaceiras e aproveitadoras. Só devíamos aceitar máquinas inteligentes autónomas se sua bússola moral for semelhante à nossa. Mas, dado que a Jurisprudência e o Direito derivados da Ética da Máquina e da Moralidade Humana estão atrasadas, tão cedo não podemos esperar uma moralidade genérica para máquinas cognitivas.

Em suma, sabemos, a partir da moral humana, que podemos ser sumamente solidários com os elementos do nosso grupo, enquanto competimos e guerreamos com os grupos rivais. Sabemos também que o domínio ético é o do *dever ser* e que, por vezes, actuamos de acordo com o que podemos, e não com o que reconhecemos ser esse tal dever. A investigação no domínio da ética, orientada para a implementação de uma IA benévola, não resolverá todas as incongruências e contradições desse domínio; mas, um melhor conhecimento da moral humana, dos fundamentos que estão para além da sua concretização em preceitos, é o único caminho para irmos mitigando os efeitos nefastos

8 Luís Moniz Pereira, The Anh Han, António Barata Lopes. Employing AI to Better Understand Our Morals. *Entropy* 24(1): 10, 2022.
Luís Moniz Pereira, António Barata Lopes. Machine Ethics: From Machine Morals to the Machinery of Morality. In: *Studies in Applied Philosophy, Epistemology and Rational Ethics* (SAPERE, volume 53), Springer Nature AG, Switzerland, 2020.
Luís Moniz Pereira, António Barata Lopes, Máquinas Éticas: Da Moral da Máquina à Maquinaria Moral, Colecção: "*Outros Horizontes*", NOVA.FCT Editorial, Campus FCT-UNL, 2829-516 Caparica, Portugal, 2020.

da competição sem regras, e para uma *inter-moralidade* agregadora de agentes cognitivos autónomos e colaborativos. Esses agentes poderão e deverão, também, competir; mas de acordo com regras explicitadas.

O fantasma da super IA (AGI) é um mito que nos desvia a atenção do importante. Para nós, o risco de extinção pela AGI não é sobre a IA se apoderar de nós, humanos, mas sobre o uso que os humanos farão da IA sem a devida preocupação coordenada com outros humanos. E essas atividades cumulativas podem resultar em efeitos colaterais incontroláveis, colocando a espécie em risco. Não vamos colocar esse problema, o verdadeiro, debaixo do tapete, levantando o fantasma de uma super IA. Os fantasmas de tudo o que traz ainda mais lucro estão em abundância entre nós. Mas agora os desenvolvedores de referência têm ferramentas de IA mais poderosas para promover seus erros, e inesperados efeitos colaterais emergentes surgirão, para além do controlo de qualquer pessoa ou nação.