DOI: 10.1590/s0103-4014.2021.35101.004

Inteligência Artificial e sociedade: avanços e riscos

Jaime Simão Sichman [I]

Introdução

A inteligência artificial (IA), surgida na década de 1950, tem sua origem praticamente confundida com a própria origem do computador. Mais precisamente, no verão de 1956, ocorreu a Darthmouth College Conference,[1] que é considerada o marco inicial da IA. Os pesquisadores reconhecidos como pais da área, como John MacCarthy, Marvin Minsky, Alan Newell e Herbert Simon, entre outros, participaram desse evento e tiveram trajetórias científicas que estabeleceram marcos nesse fascinante domínio da Computação.

Como o nome mesmo insinua, a área sempre foi cercada de enormes expectativas, e em inúmeras vezes essas não foram completamente atingidas. Desse modo, a oscilação de humor em relação à área assemelha-se a uma curva senoidal, havendo períodos de grande entusiasmo e grande financiamento (como ocorre agora) seguidos por outros de decepção e recursos escassos. Estes últimos são conhecidos como *AI Winter* (Inverno da IA), como foram por exemplo os períodos entre 1975/1980 e 1987/1993.

Atualmente, atravessamos novamente um período de euforia sobre os possíveis benefícios que a IA pode prover. Tal otimismo se justifica por uma conjunção de três fatores fundamentais: (i) o custo de processamento e de memória nunca foi tão barato; (ii) o surgimento de novos paradigmas, como as redes neurais profundas, possibilitados pelo primeiro fator e produzindo inegáveis avanços científicos; e (iii) uma quantidade de dados gigantesca disponível na internet em razão do grande uso de recursos tais como redes e mídias sociais. Tal entusiasmo, entretanto, vem sido acompanhado por uma série de temores, alguns dos quais fundados.

O objetivo deste artigo é prover informações para que o leitor comum possa melhor entender os principais aspectos da IA, em que ela difere da computação convencional e como ela pode ser inserida nos processos organizacionais da sociedade humana. Além disso, busca evidenciar os grandes avanços e potenciais riscos que essa tecnologia, tal como qualquer outra, pode provocar caso os atores envolvidos na produção, utilização e regulação de seu uso não criem um espaço de discussão adequado destas questões.

O que vem a ser IA?

Sempre que ocorre um entusiasmo com os resultados de uma tecnologia, existe uma tendência da mídia em fornecer definições e explicações, por vezes

não muito precisas, dos seus principais aspectos. Isso é, certamente, o que ocorre com a IA nos dias de hoje.

Em primeiro lugar, cabe ressaltar que não existe uma definição acadêmica, propriamente dita, do que vem a ser IA. Trata-se certamente de um ramo da ciência/engenharia da computação, e portanto visa desenvolver sistemas computacionais que solucionam problemas. Para tal, utiliza um número diverso de técnicas e modelos, dependendo dos problemas abordados. Portanto, é inadequado utilizar-se expressões como "a IA da empresa X"; mais adequado (porém com menos apelo) seria dizer "um sistema da empresa X que utiliza técnicas de IA".

Ao invés de tentar fornecer uma definição de IA, mais adequado seria tentar caracterizar quais são os objetivos da área. Uma das primeiras tentativas desta abordagem, proposta em Rich e Knight (1991), é a seguinte: o objetivo da IA é desenvolver sistemas para realizar tarefas que, no momento: (i) são mais bem realizadas por seres humanos que por máquinas, ou (ii) não possuem solução algorítmica viável pela computação convencional.

Para entender melhor essa definição, necessita-se esclarecer o que vem a ser um *algoritmo*, palavra que também é bastante citada na mídia, às vezes de modo não muito preciso. Um algoritmo nada mais é do que uma sequência finita de ações que resolve um certo problema. Uma receita culinária, como a de um risoto, é um algoritmo. Assim, um algoritmo pode resolver problemas de tipos bastante diferentes: cálculo estrutural (projeto de uma ponte), processamento de dados (geração de uma folha de pagamentos) ou planejamento (definição de um pacote de turismo).

Qual a principal diferença entre esses problemas? Basicamente, certos problemas têm soluções exatas, como o projeto da ponte, o processamento da folha de pagamentos e a receita do risoto. Solução exata, nesse caso, significa que se os passos definidos no algoritmo forem executados exatamente na ordem definida, ter-se-á ao final uma ponte que resistirá às intempéries, uma folha de pagamentos sem futuros problemas com o fisco e um delicioso risoto à moda italiana.

Por outro lado, problemas como a definição do pacote de turismo não têm uma solução exata, ou uma única solução. Outros exemplos similares são produção de diagnósticos (médicos, legais), geração automática de diálogos, reconhecimento de imagens etc. No caso do pacote de turismo, como garantir que é o melhor a ser adquirido? Deve-se escolher primeiro o voo ou o hotel? Quais datas teriam um custo menor? Existe disponibilidade nessas datas para todos os recursos desejados (hotéis, voos, passeios), e em caso positivo as férias podem ser marcadas nesse período?

Uma possível abordagem para solucionar tais problemas seria tentar gerar *as possíveis soluções* até que se obtenha a primeira delas, ou até que se encontre a melhor delas, caso existam várias soluções. Tal abordagem, apesar de teoricamente plausível, quase sempre é inviável na prática: a quantidade de possíveis soluções geradas é muito grande, e mesmo com um computador muito potente

levaria muito tempo para obtê-las. Por exemplo, um problema de definição de rotas entre cidades poderia levar centenas de dias de processamento![2]

Assim, tais problemas são usualmente mais bem solucionados por seres humanos, e na maioria dos casos de interesse não possuem solução algorítmica viável (em tempo de processamento) pela computação convencional.

Uma pergunta que se coloca então é a seguinte: Como nós, humanos, solucionamos esses problemas? Uma possível resposta é que utilizamos, de modo inato, um mecanismo de busca e poda: (i) geramos soluções candidatas ... mas quase nunca todas elas! (ii) escolhemos a melhor solução... de acordo com certo critério! e (iii) eventualmente, analisamos *a posteriori* o efeito das escolhas feitas... e as alteramos para o futuro i.e., aprendemos!

Assim, o domínio de IA se caracteriza por ser uma coleção de modelos, técnicas e tecnologias (busca, raciocínio e representação de conhecimento, mecanismos de decisão, percepção, planejamento, processamento de linguagem natural, tratamento de incertezas, aprendizado de máquina) que, isoladamente ou agrupadas, resolvem problemas de tal natureza. Para tal, podem utilizar paradigmas distintos, sendo os principais os paradigmas simbólico, conexionista, evolutivo e probabilístico.

Segundo o paradigma *simbólico*, deve-se inicialmente identificar o conhecimento do domínio (modelo do problema), para então representá-lo utilizando uma linguagem formal de representação e implementar um mecanismo de inferência para utilizar esse conhecimento.

Já no paradigma *conexionista*, a linguagem é uma rede de elementos simples, inspirada no funcionamento do cérebro, onde neurônios artificiais, conectados em rede, são capazes de aprender e de generalizar a partir de exemplos. O raciocínio consiste em aprender diretamente a função entrada-saída. Matematicamente, trata-se de uma técnica de aproximação de funções por regressão não linear.

O paradigma *evolutivo*, por sua vez, utiliza um método probabilístico de busca de soluções de problemas (otimização), onde soluções são representadas como indivíduos, aos quais se aplicam técnicas "inspiradas" na teoria da evolução como hereditariedade, mutação, seleção natural e recombinação (ou *crossing over*), para selecionar para as gerações seguintes os indivíduos mais adaptados, i.e., os que maximizam uma função objetivo (ou *fitness function*).

Finalmente, o paradigma *probabilístico* utiliza modelos para representar o conceito estatístico de independência condicional, a partir de relacionamentos causais no domínio. A inferência consiste em calcular a distribuição condicional de probabilidades dessa distribuição, e em alguns casos particulares de topologia, existem algoritmos bastante eficientes.

Agentes inteligentes

Uma contribuição muito importante foi o surgimento do conceito de *agente inteligente* (Russell; Norvig, 2010), proposto em 1995, que se tornou um paradigma integrador da área. Esse paradigma gerou uma nova área de pes-

quisa, denominada *agentes autônomos e sistemas multiagentes*, dedicada a investigar como as acima mencionadas técnicas de IA poderiam ser integradas de modo mais eficaz e efetivo em um único agente e também como um conjunto destes agentes poderia interagir de forma coordenada e cooperativa, visando resolver um problema quando nenhum deles de forma isolada poderia fazê-lo. Um conjunto de veículos autônomos seria um exemplo de um sistema multiagentes: não basta que cada um decida o melhor roteiro para atingir a meta de seu passageiro, mas é necessário que os veículos cooperem e se coordenem, para não causarem acidentes, como usualmente ocorre com condutores humanos.

Nessa nova e fascinante área de pesquisa, surgiram algumas definições importantes do que seria um agente, como a inicialmente proposta por Wooldridge (1997 apud Jennings, 1999, p.1): "Um agente é um sistema computacional encapsulado que está situado em algum ambiente, e que é capaz de ação autônoma e flexível naquele ambiente, a fim de cumprir seus objetivos".

A inserção da dimensão organizacional e a interação com os usuários foi proposta na sequência em Boissier e Sichman (2004, p.5): "Um agente é entidade real ou virtual, que é autônoma, pró-ativa, reativa e social, sendo capaz de exibir atividade organizada de modo a atingir seus objetivos, eventualmente interagindo com usuários".

Em ambas as definições, menciona-se o conceito de *autonomia*, crucial para que se possa refletir sobre os possíveis efeitos positivos e negativos da interação desses sistemas com os seres humanos.

Agentes autônomos

Costuma-se encontrar na literatura de IA várias definições para o termo autonomia. Quase todas são definições relacionais, associadas a pelo menos quatro significados muito diferentes, como inicialmente discutido em Sichman (1995, p.50):

- autonomia em relação ao *design*: um agente é autônomo se ele tem sua própria existência, independente da existência de outros agentes, como proposto em (Demazeau; Müller 1990). Anos mais tarde, tal abordagem materializou-se na chamada "computação baseada em serviços";
- autonomia em relação ao *ambiente*: um agente autônomo é um agente que deve trabalhar em ambientes dinâmicos e incertos, que só podem ser percebidos de modo imperfeito, que podem mudar como resultado de ações que não são controladas pelo próprio agente e sobre o quais os efeitos de suas ações nem sempre são previsíveis. Essa definição de autonomia é normalmente encontrada nos trabalhos de robótica móvel, como os elencados por (Nilsson, 1994) e também está muito próximo da noção de autopoiese citada em (Bourgine, 1995);
- autonomia em relação aos *próprios objetivos*: um agente autônomo é um agente que pode atingir seus objetivos por conta própria. Ele não tem não necessidade *a priori* de cooperar com outros agentes, e caso decida

fazê-lo tal escolha se deve a uma possível melhoria de sua atuação. Essa noção de autonomia, bem como outra em relação à localidade (local ou global) da tarefa a ser realizada, foi usada em (Demazeau; Müller, 1990) para classificar comportamentos possíveis de um agente de negócio (coabitação, cooperação, colaboração e distribuição), sendo posteriormente refinada em (Ferber, 1995) para propor a noção de situações de interação entre agentes;

- autonomia em relação às *motivações*: um agente autônomo é um agente que tem a liberdade de escolha para interagir socialmente. É a partir do conteúdo de seu estado mental que ele decide cooperar ou não, adotar objetivos de outros agentes ou não etc., como discutido em (Castelfranchi, 1990). Em outras palavras, o significado do termo significa que um agente não é necessariamente benevolente, podendo *decidir* atingir um objetivo ou não, cooperar com outros agentes ou não.

Mais recentemente, tal caráter plurifacetado de noção de autonomia foi reiterado em (Dignum, 2019, p.18):

> É importante notar que sistema autônomo é um termo impróprio, pois nenhum sistema é autônomo em todas as situações e para todas as tarefas [...] Autonomia não é uma propriedade intrínseca de um sistema, mas sim o resultado da sua interação com a tarefa, contexto e ambiente [...] Não se trata de uma propriedade emergente, mas de algo que deve ser projetado no sistema.

Certamente, a definição de autonomia em relação às suas motivações é a que provoca mais discussões no contexto das atuais e (potencialmente) futuras aplicações de IA. Deve-se certamente discutir até que ponto se deseja que um dispositivo inteligente seja autônomo nesse sentido: talvez seja adequado aceitar a autonomia de um aspirador de pó robótico (afinal, não seria necessário informá-lo qual local deve ser limpo em primeiro lugar), mas talvez esse não fosse o caso de um agente inteligente de reserva de viagens (talvez fosse mais adequado que ele sugerisse opções mas não tomasse a iniciativa de comprá-las antes de uma confirmação do usuário).

Um trabalho muito interessante que propõe uma discussão nesse sentido é o proposto em Falcone e Castelfranchi (2000), onde se discutem graus de autonomia distintos que podem ser outorgados a esses agentes artificiais, fundeados em métricas de confiança baseadas no histórico de interações anteriores. Similarmente ao que ocorre na sociedade humana, talvez numa primeira interação entre um docente e seu orientado, o primeiro explique muito mais detalhadamente os procedimentos experimentais que devem ser realizados; à medida que mais interações bem-sucedidas ocorram, no futuro provavelmente pode ocorrer que o docente delegue certa autonomia de planejamento ao seu orientado. Um exemplo de autonomia de planejamento, no contexto de interações entre agentes inteligentes autônomos, pode ser visto em Maia e Sichman (2020).

No caso particular de interações entre tais agentes inteligentes e humanos, um grande desafio é incorporar tais graus de autonomia nos chamados sistemas sociotécnicos.

Interação humano-agente

Antes de analisarmos os avanços e riscos potenciais da IA *per se*, cabe introduzir o conceito de Sistemas Sociotécnicos (SST). O termo foi cunhado por Eric Trist, Ken Bamforth e Fred Emery, na era da Segunda Guerra Mundial, derivado de seu estudo com trabalhadores em minas de carvão inglesas no Instituto Tavistock em Londres (Trist et al., 2013).

A abordagem, segundo Appelbaum (1997), parte da premissa de que organizações são compostas de elementos sociais e técnicos, que trabalham conjuntamente para realizar as tarefas organizacionais. Tal atuação conjunta gera tanto produtos físicos como resultados sociais/psicológicos. O foco da abordagem consiste em possibilitar que os dois elementos gerem resultados positivos, diferentemente dos métodos convencionais em que as pessoas se adaptem e se ajustem aos elementos técnicos.

Tais sistemas já estão presentes em nossas vidas há pelo menos duas décadas: basta pensar nas nossas experiências com diversos tipos de *call-center* ou serviços bancários. Atualmente, na maioria dos casos, os elementos técnicos fornecem subsídios para que humanos possam tomar decisões. Há instâncias para recursos que podem, em certos casos, alterar decisões tomadas de forma equivocada, inclusive aplicando eventualmente sanções aos atores envolvidos para aprimorar os resultados futuros do sistema. Entretanto, a inserção da tecnologia de IA em tais sistemas pode alterar tal prática, fazendo que os próprios elementos técnicos possam tomar algumas decisões. Tal mudança de paradigma não é necessariamente boa ou ruim, mas tais sistemas necessitam incorporar outras propriedades inerentes à interação humana.

IA responsável

Em seu trabalho seminal sobre IA Responsável, Virginia Dignum (2019) sintetiza num livro fascinante como deve-se desenvolver e utilizar IA de modo responsável. A autora advoga que uma postura ética deve ser adotada em três instâncias distintas:

- no processo de *projeto* de tais sistemas, garantindo que as equipes tenham em mente e antevejam as possíveis consequências do sistema para os indivíduos e sociedades;
- no projeto do *comportamento* de tais sistemas, visando representar de forma adequada capacidades de raciocínio ético nos agentes inteligentes;
- no *código de conduta* dos projetistas e desenvolvedores, mediante uma regulação adequada e processos de certificação que garantam um comportamento adequado dos atores envolvidos, como já existe em outras profissões.

Para a primeira dimensão (ética no projeto), a autora propõe uma abordagem denominada *ART of AI*, que garante que os valores humanos e princípios éticos, suas prioridades e escolhas sejam explicitamente incluídos nos processos de *design* de forma transparente e sistemática. Tal abordagem é composta por três partes:

- prestação de contas (*accountability*) se refere à necessidade do sistema de IA explicar e justificar suas decisões e ações para seus parceiros, usuários e outros com quem o sistema interage;
- responsabilidade (*responsibility*) se refere ao papel das próprias pessoas e à capacidade dos sistemas de IA de responder por uma decisão e identificar erros ou resultados inesperados. À medida que a cadeia de responsabilidade cresce, são necessários meios para vincular as decisões dos sistemas de IA ao uso justo dos dados e às ações das partes interessadas envolvidas na decisão do sistema;
- transparência (*transparency*) refere-se à necessidade de descrever, inspecionar e reproduzir os mecanismos pelos quais os sistemas de IA tomam decisões e aprendem a se adaptar ao seu ambiente e à governança dos dados utilizados e criados. Os algoritmos de IA atuais são basicamente caixas-pretas. No entanto, reguladores e usuários exigem explicação e clareza sobre os dados usados. Métodos são necessários para inspecionar algoritmos e seus resultados e para gerenciar dados, sua proveniência e sua dinâmica.

Quanto à segunda dimensão (ética no comportamento), deve-se levar em conta que as sociedades humanas usualmente seguem *normas* para facilitar a interação. Tais normas, em muitos casos, levam em conta valores morais para embasar decisões. Assim, um grande desafio é incorporar tais normas e valores em sistemas de IA. Tal assunto vem sendo tratado pelos pesquisadores da área há mais de vinte anos, por exemplo na série de workshops denominada Coordination, Organization, Institutions and Norms in agent systems (Coin, 2005). Trata-se de embasar tais agentes autônomos com mecanismos de decisão que possam ser também baseados em sentimentos e valores morais, como proposto em Bazzan et al. (2002), ou que possam julgar a dimensão ética de seu próprio comportamento e dos comportamentos de outros agentes, como apresentado em Cointe et al. (2016). Além dessa perspectiva individual, necessita-se também prover mecanismos de governança adequados, que possam eventualmente sancionar comportamentos distintos dos esperados por estes agentes, como proposto em Nardin et al. (2016).

A questão da transparência é uma condição necessária para tais agentes inteligentes possam argumentar e explicar as decisões por eles tomadas.

IA explicável e IA para o bem

Os chamados sistemas de IA explicáveis incorporam processos de explicação que permitem aos usuários obter informações sobre os modelos e decisões do

sistema. O Explainable Artificial Intelligence workshop (XAI, 2018), evento satélite da ECAI/IJCAI 2018, ocorrida em Estocolmo, Suécia, possibilitou reunir pesquisadores interessados em IA, interação homem-computador, modelagem cognitiva e teorias cognitivas de explicação e transparência. Um tema fundamental, dado seu sucesso recente, foi como adicionar explicações a técnicas de aprendizado profundo, quase sempre baseados em modelos de caixa-preta, cujos parâmetros internos e seus respectivos valores são pouco entendidos pelo usuário.

A preocupação com as finalidades de uso de sistemas de IA também têm sido objeto de debate nos últimos anos. O AI for Social Good workshop (AI4G 2019), evento satélite da IJCAI 2019, ocorrida em Macao, China, teve como objetivo explorar como a IA poderia contribuir para resolver problemas sociais.

Já o Responsible Artificial Intelligence Agents workshop (Raia, 2019), evento satélite do AAMAS 2019, ocorrido em Montreal, Canadá, reuniu pesquisadores de IA, ética, filosofia, robótica, psicologia, antropologia, ciências cognitivas, direito, estudos de governança regulatória e engenharia para discutir e trabalhar sobre os complexos desafios relacionados ao projeto e à regulamentação de sistemas de IA. Concentrou-se em três aspectos que juntos podem garantir que a IA seja desenvolvida para o bem da sociedade (por exemplo, contribuindo para os objetivos de desenvolvimento sustentável da ONU), usando processos verificáveis e responsáveis, e que seu impacto seja governado por instituições e mecanismos justos e inclusivos.

Tais preocupações também têm norteado a criação de centros interdisciplinares para a formação de alunos na área. A UK Research and Innovation (UKRI) é uma agência de financiamento britânica que trabalha em parceria com universidades, organizações de pesquisa, empresas, instituições de caridade e governo para criar o melhor ambiente possível para a pesquisa e inovação florescerem. Em particular, apoiou recentemente a criação de 16 Centros de Treinamento de Doutorado (CDT) em Inteligência Artificial, visando formar 1.000 estudantes de doutorado para explorar o potencial da IA para transformar a maneira como trabalhamos e vivemos. As áreas de pesquisa são diversas, envolvendo desde saúde e mudanças climáticas a ética e música. Entre tais centros, podem-se destacar o Centre for Doctoral Training in Safe & Trusted AI (STAI, 2019), envolvendo o King's College e o Imperial College, em Londres, e o Centre for Doctoral Training in Accountable, Responsible and Transparent AI (ART-AI, 2019), sediado na University of Bath.

Avanços e riscos da IA

Há cinco anos, num artigo divulgado no *Jornal da USP* que foi escrito juntamente com meus colegas Fabio Cozman e Claudio Pinhanez, ambos hoje à frente do Centro de Inteligência Artificial da USP (C4AI),[3] já apontávamos para os grandes avanços da IA nas últimas décadas (Sichman et al., 2016):

> [...] é inegável o tremendo sucesso pragmático de tecnologias ligadas à IA. Sistemas de busca de informação e de recomendação de produtos são

> parte de nossa experiência cotidiana. Tais produtos aprendem a partir de dados e decidem com base em regras e em experiências passadas. O sistema financeiro também depende fortemente de programas com capacidade de raciocínio e decisão, que hoje comandam grandes investimentos em bolsas ao redor do mundo. Usamos hoje também sistemas de diagnóstico automático, sistemas comerciais de análise e organização de documentos e até mesmo veículos aéreos não tripulados (drones) para fins pacíficos e militares. Em resumo, nosso mundo já é um mundo no qual máquinas apresentam comportamentos tipicamente associados à "inteligência" [...].

Nesse mesmo artigo, também mostramos que os temores a respeito de robôs aniquiladores da raça humana não poderiam ser construídos com a tecnologia atual:

> [...] Considerando a tecnologia de computadores em silício, base da computação hoje e nas próximas décadas, é difícil imaginar como isso seria possível. Um computador com uma capacidade de processamento equivalente àquela de um cérebro humano teria de ser pelo menos mil vezes mais rápido que o mais rápido computador hoje existente. E teria que consumir energia na grandeza de hidrelétricas e dissipar calor usando o sistema de ar-condicionado de um arranha-céu. Como produzir – e dissipar! – tanta energia em um robô é um desafio que a tecnologia do silício dificilmente conseguirá resolver. (ibidem)

Num artigo interessante, Thomas Dietterich e Eric Horvitz (2015) elencaram cinco classes de riscos envolvendo o uso de sistemas de IA:

- *falhas (bugs)*: Quaisquer sistemas de software apresentam falhas. Vários sistemas de software convencionais foram desenvolvidos e validados para atingir altos níveis de garantia de qualidade; por exemplo, sistemas de piloto automático e de controle de espaçonaves são cuidadosamente testados e validados. Práticas semelhantes devem ser aplicadas aos sistemas de IA;
- *segurança (cybersecurity)*: Os sistemas de IA são tão vulneráveis quanto qualquer outro software a ataques cibernéticos. Por exemplo, ao manipular dados de treinamento ou preferências e trade-offs codificados em modelos de utilidade, adversários podem alterar o comportamento desses sistemas;
- *aprendiz de feiticeiro (sorcerer's apprentice)*: Um aspecto importante de qualquer sistema de IA que interage com as pessoas é que ele deve raciocinar sobre o que estas pretendem, em vez de executar comandos literalmente. Um sistema de IA deve analisar e compreender se o comportamento que um ser humano está solicitando pode ser julgado como "normal" ou "razoável" pela maioria das pessoas;
- *autonomia compartilhada (Shared autonomy)*: Construir esses sistemas colaborativos levanta um quarto conjunto de riscos decorrentes de desafios sobre fluidez de engajamento e clareza sobre estados internos e

objetivos dos envolvidos no sistema. Criar sistemas em tempo real onde o controle precisa mudar rapidamente entre as pessoas e os sistemas de IA é difícil;

- *impactos socioeconômicos*: Precisamos entender as influências da IA na distribuição de empregos e na economia de forma mais ampla. Essas questões perpassam a ciência e engenharia da computação, chegando ao domínio das políticas e programas econômicos que podem garantir que os benefícios dos aumentos de produtividade baseados em IA sejam amplamente compartilhados.

Dentre estes riscos, os três últimos merecem maior atenção, por serem mais particulares ao uso da tecnologia de IA.

Conclusões

Sob qualquer perspectiva e métrica, é inegável que a IA alcançou um tremendo sucesso. As maiores empresas da economia mundial, como as Big Techs, são efetivamente empresas de IA. Como mencionado na introdução, tal sucesso se deu pelo barateamento dos custos de processamento e de memória, surgimento de novos paradigmas, como as redes neurais profundas e a enorme quantidade de dados disponível nas redes e mídias sociais.

Novamente fazendo referência ao trabalho de Virginia Dignum (2019) sobre IA Responsável, a questão ética dos sistemas de IA que já fazem parte do nosso quotidiano deve ser ressaltado:

> O desenvolvimento e o uso da IA levantam questões éticas fundamentais para a sociedade, que são de vital importância para o nosso futuro. Já existe muito debate sobre o impacto da IA no trabalho, interações sociais (incluindo cuidados de saúde), privacidade, justiça e segurança (incluindo iniciativas de paz e guerra). O impacto social e ético da IA abrange muitos domínios, por exemplo, os sistemas de classificação de máquinas levantam questões sobre privacidade e preconceitos e veículos autônomos levantam questões sobre segurança e responsabilidade. Pesquisadores, decisores políticos, indústria e sociedade reconhecem a necessidade de abordagens que garantam as tecnologias de IA de uso seguro, benéfico e justo, para considerar as implicações da tomada de decisão ética e legalmente relevante pelas máquinas e o status ético e legal da IA. Essas abordagens incluem o desenvolvimento de métodos e ferramentas, atividades de consulta e treinamento e esforços de governança e regulamentação.

Para encerrar, cabe relembrar uma frase do fundador da Cibernética, Norbert Wiener, que faz parte do artigo "Some Moral and Technical Consequences of Automation", publicado na revista *Science*, em 1960: "*Se usarmos, para atingir nossos objetivos, um órgão mecânico em cujo funcionamento não podemos interferir de forma eficaz ... é melhor estarmos bem certos de que o propósito colocado na máquina é aquele que realmente desejamos*".

Notas

1 Disponível em: <https://250.dartmouth.edu/highlights/artificial-intelligence-ai-coined-dartmouth>.

2 Imaginando-se um método de busca em largura, que para atingir a cidade destino deva avaliar roteiros que passam por oito cidades intermediárias, onde cada uma delas tivesse ligações diretas com dez outras cidades e que o computador pudesse analisar mil alternativas por segundo.

3 Disponível em: <http://c4ai.inova.usp.br/>.

Referências

AI4G. AI for social good, IJCAI 2019 Workshop, Macao, China. Disponível em: <https:// aiforgood2019.github.io/>.

APPELBAUM, S. H. Socio-technical systems theory: an intervention strategy for organizational development. *Management Decision*, v.35, n.6, 1997.

ART-AI. Centre for Doctoral Training in Accountable, Responsible and Transparent AI, University of Bath, UK, 2019. Disponível em: <https://cdt-art-ai.ac.uk/>.

BAZZAN, A. L. et al. Evolution of agents with moral sentiments in an iterated prisoner's dilemma exercise. In: *Game theory and decision theory in agent-based systems.* Springer. 2002. p.43-64.

BOISSIER, O.; SICHMAN, J. Organization oriented programming. Tutorial Notes. In: 3rd. INTERNATIONAL CONFERENCE ON AUTONOMOUS AGENTS AND MULTI-AGENT SYSTEMS (AAMAS 2004), New York, USA, 2004.

BOURGINE, P. Models of self-teaching agents and the emergence of the symbolic level. In: *Pre-proceedings of the invited lectures of the 1st European Conference on Cognitive Science*, St. Malo, 1995.

CASTELFRANCHI, C. Social power: A point missed in multi-agent, DAI and HCI. In: DEMAZEAU, Y.; MÜLLER, J.-P. (Ed.) *Decentralized A. I.* Amsterdam: Elsevier Science Publishers B. V., 1990. p.49-62.

COIN. Coordination, Organization, Institutions and Norms in Agent Systems (COIN). 2005. The International Workshop Series. Disponível em: <http://www2.pcs.usp.br/coin/>.

COINTE, N. et al. Ethical judgment of agents' behaviors in multi-agent systems. In: JONKER, C. M. et al. (Ed.) *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems.* Singapore, May 9-13, 2016, p.1106-14. ACM.

DEMAZEAU, Y.; MÜLLER, J.-P. Decentralized artificial intelligence. In: DEMAZEAU, Y.; MÜLLER, J.-P. (Ed.) *Decentralized A. I.* Amsterdam: Elsevier Science Publishers B. V., 1990. p.3-13.

DIETTERICH, T. G.; HORVITZ, E. Rise of concerns about AI: reflections and directions. *Communications of the ACM*, v.58, n.10, p.38-40, 2015.

DIGNUM, V. *Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way.* Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, 2019.

FALCONE, R.; CASTELFRANCHI, C. Grounding autonomy adjustment on delegation and trust theory. *Journal of Experimental & Theoretical Artificial Intelligence*, v.12, n.2, p.149-51, 2000.

FERBER, J. *Les Systèmes Multi-Agents*: Vers une Intelligence Collective. Paris: InterEditions, 1995.

JENNINGS, N. R. Agent-oriented software engineering. In: GARIJO, F. J.; BOMAN, M. (Ed.) *Multiagent System Engineering, 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW '99*. Valencia, Spain, June 30 - July 2, 1999, *Proceedings*, v.1647 of Lecture Notes in Computer Science, p.1-7.

MAIA, A. V.; SICHMAN, J. S. Representing planning autonomy in agent organizational models. *Theoretical Computer Science*, v.805, p.92-108, 2020.

NARDIN, L. G. et al. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *Knowledge Engineering Review*, v.31, n.2, p.142-66, 2016.

NILSSON, N. J. Teleo-reactive programs for agent control. *Journal of Artificial Intelligence*, v.1, p.139-58, 1994.

RAIA. Responsible artificial intelligence agents, AAMAS 2019 Workshop, Montreal, Canada. Disponível em: <https://raia2019.blogs.dsv.su.se/>.

RICH, E.; KNIGHT, K. *Artificial intelligence*. 2.ed. s.l.: McGraw-Hill, 1991.

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence - A Modern Approach, Third International Edition*. s.l.: Pearson Education, 2010.

SICHMAN, J. S. *Du Raisonnement Social Chez les Agents*: Une Approche Fondée sur la Théorie de la Dépendance. Grenoble, 1995. Thèse (Doctorat) – Institut National Polytechnique de Grenoble.

SICHMAN, J. S. et al. É possível a máquina superar o ser humano? *Jornal da USP*, n.XXX1, 2016.

STAI. Centre for Doctoral Training in Safe & Trusted AI, King's College, UK. 2019. Disponível em: <https://www.kcl.ac.uk/informatics/research/ safe-trusted-ai>.

TRIST, E. L. et al. *Organizational Choice (RLE: Organizations)*: Capabilities of Groups at the Coal Face Under Changing Technologies. s.l.: Routledge, 2013.

WOOLDRIDGE, M. J. Agent-based software engineering. *IEE Proceedings on Software Engineering*, v.144, n.1, p.26-37, 1997.

XAI. Explainable artificial intelligence, ECAI/IJCAI 2018 Workshop, Stockholm, Sweden. Disponível em: < http://home.earthlink.net/ ~dwaha/research/meetings/ faim18-xai/>.

RESUMO – Este artigo tem como objetivo prover informações para que o leitor comum possa melhor entender os principais aspectos da IA, em que ela difere da computação convencional e como ela pode ser inserida nos processos organizacionais da sociedade humana. Além disso, busca evidenciar os grandes avanços e potenciais riscos que essa tecnologia, tal como qualquer outra, pode provocar caso os atores envolvidos na sua

produção, utilização e regulação não criem um espaço de discussão adequado destas questões.

PALAVRAS-CHAVE: Inteligência Artificial, Agentes inteligentes, Sistemas multiagentes, Agentes normativos, Regulação de sistemas autônomos.

ABSTRACT – The goal of this article is to provide information for the common reader to understand better the main aspects of AI, how it differs from conventional computing and how it can be inserted in the organizational processes of human society. The article also seeks to highlight the great advances and potential risks of this technology, like of any other, if the actors involved in its production, use and regulation do not create adequate space for discussing these issues.

KEYWORDS: Artificial intelligence, Intelligent agents, Multi-agent systems, Normative agents, Regulation of autonomous systems.

Jaime Simão Sichman é doutor em Engenharia de Computação pelo Institut National Polytechnique de Grenoble (INPG), França. É professor titular do Departamento de Engenharia de Computação e Sistemas Digitais (PCS) da Escola Politécnica (EP) da Universidade de São Paulo (USP), onde ocupa os cargos de Chefe de Departamento (PCS) e de Presidente da Comissão de Pesquisa (EP). @ – jaime.sichman@usp.br / https://orcid.org/0000-0001-8924-9643.

Recebido em 10.3.2021 e aceito em 12.3.2021.

[I] Universidade de São Paulo, Escola Politécnica, Departamento de Engenharia de Computação e Sistemas Digitais, São Paulo, Brasil.