Inteligência Artificial

DOI: 10.1590/s0103-4014.2021.35101.002

No canal da Inteligência Artificial – Nova temporada de desgrenhados e empertigados

Fabio Gagliardi Cozman [1]

Desgrenhados e empertigados

A construção de inteligências artificiais sempre esteve cercada de controvérsias, não apenas sobre seus limites, mas também sobre quais os objetivos a perseguir. Parece haver dois estilos fundamentalmente diferentes de abordagem em Inteligência Artificial (IA): de um lado, um estilo empírico, fortemente respaldado por observações sobre a biologia e psicologia de seres vivos, e pronto para abraçar arquiteturas complicadas que emergem da interação de muitos módulos díspares; de outro lado, um estilo analítico e sustentado por princípios gerais e organizadores, interessado em concepções abstratas da inteligência e apoiado em argumentos matemáticos e lógicos. Por volta de 1980 os termos *scruffy* e *neat* foram cunhados para se referir respectivamente a esses dois estilos de trabalho. Robert P. Abelson (1981) apresentou aparentemente a primeira publicação que discute os dois termos, atribuindo a distinção a um colega não nomeado, mas, segundo Abelson, facilmente identificável – que, segundo a literatura, deve ser Roger Schank (Nilsson, 2009).

Os termos *scruffy* e *neat* não são exatamente elogiosos; de certa forma, cada um desses termos é especialmente adequado para ser usado por um time contra o outro. Os *scruffies* são desgrenhados, perdidos em sistemas de confusa complexidade. Os *neats* são empertigados, teorizando em torres de marfim e desconectados dos detalhes do mundo real. Abelson identifica essas atitudes gerais em muitas atividades humanas, em arte, em política, em ciência. Há pessoas que favorecem resultados obtidos por meio de experimentos, não se importando se soluções fogem de rotinas preestabelecidas, enquanto outras pessoas buscam ordem e harmonia em teoria amplas. Argumentos similares podem ser encontrados em outras áreas: por exemplo, Sergiovanni (2007) discute as perspectivas *scruffy* e *neat* no ensino de administração, conectando cada um deles a uma forma de entender a profissão.

Ao longo das últimas décadas a distinção entre desgrenhados e empertigados tem sido repetidamente discutida na literatura de IA, nem sempre de maneira uniforme. Às vezes a primeira posição é apenas identificada com a disposição de lidar com sistemas de grande complexidade; nesse sentido a IA é quase inevitavelmente desgrenhada, pois ninguém imagina que a inteligência

seja produzida de forma simples (até onde sabemos, o cérebro é um órgão de extrema complexidade e resultado de um longo processo evolucionário). E a posição empertigada às vezes é reduzida à organização da disciplina para efeito de ensino e classificação de temas; tais atividades são imprescindíveis, mas não capturam a perspectiva empertigada. Neste artigo os dois estilos são tomados como visões diferentes do que se pretende em IA: os desgrenhados procuram reproduzir, sem necessariamente uma base teórica única, aspectos empíricos da inteligência observada no mundo real; os empertigados procuram construir arquiteturas que partam de princípios gerais e demonstrem comportamento inteligente, quiçá mais racional do que vemos no mundo real. O objetivo aqui é compreender como essas duas visões progrediram ao longo das últimas décadas e, principalmente, como elas se apresentam hoje e como influenciam o presente e futuro da IA.

Temporadas anteriores: de 1950 a 2010

A tensão entre desgrenhados e empertigados tem raízes históricas em IA, colocando alguns dos fundadores da disciplina em posições opostas. Em muitos momentos é difícil classificar uma determinada técnica em um campo ou outro; as próprias opiniões se confundem ao longo do tempo. Por isso, vale a pena examinar a história da IA e os debates entre desgrenhados e empertigados.

O nome Artificial Intelligence foi cunhado em 1956, tendo sido usado no título de um encontro de pesquisadores no Dartmouth College, Estados Unidos. Entre os organizadores desse encontro estavam John McCarthy, defensor de teorias lógicas e abstratas, e Marvin Minsky, adepto de arquiteturas que emergem a partir de questões empíricas. Note-se que tanto Minsky quanto McCarthy receberam, respectivamente, em 1969 e 1971, o Turing Award, o mais importante prêmio em ciência da computação. No texto de recebimento do prêmio, Minsky escreve (1970): "Uma excessiva preocupação com formalismo está impedindo o desenvolvimento da ciência de computação".[1] Por outro lado, o título da palestra correspondente de McCarthy é "Generalidade em inteligência artificial".[2]

Desde seu início a IA tem tido uma conexão forte com o estudo de lógica matemática. Essa conexão é um exemplo fundamental da abordagem empertigada; por isso, vale a pena examiná-la em mais detalhes.

Provavelmente o primeiro artigo defendendo uma IA baseada em lógica foi produzido por McCarthy em 1958; como coloca McCarthy (1958), o "artigo discute programas que manipulam, em uma linguagem formal adequada (provavelmente uma parte do cálculo de predicados), sentenças comuns instrumentais".[3] McCarthy apresenta as vantagens de representações declarativas expressas em um sistema lógico, descrevendo um programa chamado *advice taker* que receberia sentenças em lógica de primeira-ordem e seria capaz de realizar deduções lógicas. McCarthy exemplifica suas intenções com um exemplo envolvendo locais e movimentos. Por exemplo, o *advice taker* poderia ser informado que

at(*desk*, *home*),

indicando que *desk* está em *home*. Ou o *advice taker* poderia ser informado que a relação *at* satisfaz propriedades de transitividade: para todos os objetos x, y, z,

at(x,y), *at*(y,z) → *at*(x,z),

onde o símbolo → indica implicação (intuitivamente: se x está em y e y está em z, então x está em z).

A proposta do *advice taker* enfatiza a necessidade, para um artefato inteligente, de representar com eficiência o conhecimento adquirido previamente. Dada a importância de representação de conhecimento em IA (Davis et al., 1993), as propostas feitas por McCarthy em 1958 encontraram terreno favorável; lógica se tornou uma das ferramentas mais ilustres no arsenal da IA. Como exemplo, o estudo de programação baseada em lógica recebeu enorme atenção – ao ponto de o Japão centrar, por volta de 1982, boa parte dos esforços do seu Computador de Quinta-Geração em programação lógica. Os empertigados tiveram grandes vitórias.

Ao mesmo tempo, muitos programas codificados na década de 1960 e 1970 não se prendiam às rígidas convenções impostas por linguagens formais. Muitos sistemas focavam em uma tarefa específica, procurando demonstrar que um programa poderia resolvê-la, e declarando vitória quando isso acontecia – sem muita análise sobre princípios gerais de projeto que poderiam ser extrapolados para outras situações.[4] Enquanto os empertigados trabalhavam em teorias lógicas e suas potenciais aplicações, desgrenhados procuravam construir programas com variados esquemas de raciocínio e tomada de decisão. Ocasionalmente, desgrenhados duvidavam do valor de formalismos lógicos. O calor do debate entre defensores da lógica e seus críticos pode ser capturado em um influente artigo de Patrick Hayes publicado em 1977, no qual o autor defende lógica como a mais bem-sucedida linguagem desenvolvida para expressar pensamentos e inferências humanas, argumentando que muitas propostas na literatura não apresentariam a solidez semântica da lógica formal (Hayes, 1977).

Em 1983, Nils Nilsson, então presidente da American Association for Artificial Intelligence, ofereceu uma perspectiva ampla para a IA em seu Presidential Address (Nilsson, 1983). A palestra de Nilsson foi proferida após uma fase de relativo desapontamento com os resultados da pesquisa em IA, às vezes chamada do "inverno da IA". Seu tom é de otimismo, apresentando uma lista de vitórias da IA que inclui representações declarativas, formalizações do senso comum relacionado a processos físicos por meios qualitativos, teorias de linguagem baseadas em estados cognitivos. Nilsson explicitamente discute, como um desafio a ser enfrentado, a diferença de estilos entre desgrenhados e empertigados: enquanto alguns pesquisadores tomavam IA como uma arte empírica, outros a consideravam uma disciplina teórica.

Em sua análise da tensão desgrenhado/empertigado, Nilsson adota uma posição diplomática, argumentando que uma área dinâmica exige tanto pessoas que expandem a fronteira de conhecimento, sem serem inibidas por dogmas existentes, quanto pessoas que esclarecem e codificam esta fronteira. Em certa medida, essa visão se aplica a qualquer campo de pesquisa. Porém, a divergência entre desgrenhados e empertigados é mais fundamental em IA, dizendo respeito ao foco da área: seria esse foco a procura por princípios gerais que devem reger inteligências artificiais, ou seria a reprodução de comportamentos e estruturas observadas empiricamente?

Tensões similares podem ser encontradas em discussões sobre processamento computacional de linguagem natural. Princípios linguísticos tiveram grande destaque no processamento de linguagem natural durante as primeiras décadas da IA. Em particular, a década de 1980 testemunhou um grande interesse em teorias de linguagem baseadas em lógica, focando-se em vários tipos de regras gramaticais. Entretanto, diferentes soluções baseadas em gramáticas e semânticas não atingiram um sucesso significativo naquele período. Embora programas mostrassem razoável desempenho na construção de árvores sintáticas e outras estruturas, havia muita dificuldade de realmente produzir um sistema com desempenho robusto e confiável. Além das dificuldades práticas, surgiram também críticas ao esquema formal de análise de linguagem, que podem ser ilustradas em uma sentença de Abelson, na qual afirma que "compreender a cognição por meio de uma análise formal da linguagem me parece como tentar entender beisebol pela análise da física do que acontece quando um bastão idealizado atinge uma bola de beisebol idealizada"[5] (Abelson, 1981). O foco em análise linguística começou a esmaecer durante a década de 1990, quando modelos estatísticos passaram a capturar a probabilidade de termos emitidos por uma fonte a partir de termos previamente emitidos. Um exemplo importante foi o sistema *Candide*, da IBM, que traduzia textos de francês para inglês usando probabilidades de emissão de um termo com base em dois termos emitidos anteriormente (Berger et al., 1994). Desde então o processamento computacional de linguagem natural deixou de focar em modelos gramaticais. De certa forma, um novo princípio organizador foi adotado: para processar texto, coletam-se grandes quantidades de texto e estima-se um modelo estatístico de quais termos aparecem em variadas situações.

A década de 1980 testemunhou um grande interesse em IA, inclusive do ponto de vista de empreendedores e investidores. Boa parte desse interesse pode ser atribuída ao desenvolvimento de sistemas especialistas, a maioria dos quais baseada em regras capazes de codificar o conhecimento de um domínio escolhido. Linguagens formais, em particular a lógica, se solidificaram como um núcleo básico de IA. Mas ao mesmo tempo, a ampliação do leque de aplicações práticas e as dificuldades enfrentadas na manipulação de incertezas gerou muitas técnicas díspares em torno desse núcleo básico. Também é notável o trabalho feito nesse período sobre arquiteturas computacionais de suporte a atividades cognitivas; por

exemplo, Marvin Minsky (1986) continuou sua exploração de arquiteturas desgrenhadas baseadas em "Sociedade de Mentes", enquanto Allen Newell (1990) propôs uma abordagem mais estruturada na arquitetura Soar.

Merecem menção dois outros movimentos da década de 1980, um decididamente desgrenhado. Em robótica, Rodney Brooks (1990) teve grande impacto como um enfático defensor de arquiteturas pouco formalizadas e reativas. Outro grande impacto foi causado pelo reaparecimento de redes neurais, agora multicamadas e acompanhadas pelo famoso algoritmo de *backpropagation* (Rumelhart et al., 1986). Comentando sobre o debate entre desgrenhados e empertigados, Aaron Sloman (1990, p.2) vê certa dificuldade em situar as redes neurais:

> Aqui temos um ramo da IA [...] que é pesadamente matemática e entretanto, embora os princípios gerais pelos quais uma particular rede aprende durante seu período de treino sejam bem compreendidos, a operação [...] geralmente depende de uma rede de conexões totalmente opaca entre unidades de processamento e pesos que modificam suas influências entre si.[6]

Vale a pena, para indicar a situação da IA no início da década de 1990, verificar o que significava um curso introdutório em IA naquele momento. A área contava com alguns livros-texto; provavelmente o mais popular era então o livro de Elaine Rich e Kevin Knight (1991). Certamente sintonizado com o melhor da pesquisa da época, o livro reflete uma disciplina imatura e dividida entre um sem-número de técnicas rivais.[7] Um estudante dificilmente conseguia perceber como usar as técnicas apresentadas para resolver problemas reais. O livro dá grande ênfase à lógica e suas aplicações, por exemplo, em planejamento de tarefas, valorizando assim um núcleo empertigado para toda a disciplina, ao mesmo tempo em que fornece um panorama geral totalmente desgrenhado e pouco conectado com aplicações verdadeiras.

A década de 1990 não foi tão gentil para a IA; alguns se referem a esse período como um segundo "inverno" em razão das controvérsias e quedas em financiamento e interesse. Porém, foi um tempo de grandes avanços conceituais. Embora esses avanços sejam multifacetados, um grande eixo condutor foi a adoção de formalismos baseados em probabilidades, estatística e otimização. A teoria de probabilidades, que fora bastante criticada durante décadas anteriores na literatura de IA (Cheeseman, 1985), recebeu uma nova roupagem através do trabalho liderado por Judea Pearl (1988). Gradualmente métodos baseados em probabilidades e estatística passaram a ser utilizados em todos os cantos da IA.

Assim, um novo princípio norteador foi incorporado à IA, dando-se ênfase a uma racionalidade axiomatizada: o agente inteligente deve maximizar a utilidade de suas ações, medindo essa utilidade de forma numérica e ponderando a incerteza em eventos por meio de probabilidades. A IA que emergiu desses esforços dependia mais de cálculos numéricos do que de fórmulas lógicas; mais de otimização do que de conjuntos de regras. De qualquer forma, um estilo baseado em princípios unificadores.

Esse espírito pode ser sentido no livro-texto de Russell e Norvig (1995), publicado em 1995 e ainda hoje o mais popular material, em suas várias edições, para estudo da IA. O livro claramente se fundamenta num conjunto pequeno de linhas mestras, dando ênfase a técnicas para construção de agentes que se comportam de forma racional (próximos do *homo economicus* idealizado). Obviamente, é de esperar que um livro-texto bem escrito valorize os aspectos estruturantes e gerais de qualquer disciplina, procurando dar a seus leitores estudantes as mais básicas ferramentas, mesmo que a prática seja bem mais caótica. Porém, como já vimos, em IA o desejo de organizar a disciplina em torno de alguns princípios teóricos diz muito sobre o objetivo geral da empreitada; uma perspectiva empertigada dá ênfase a artefatos que sigam diretrizes racionais de forma útil, em detrimento de arquiteturas que emergem de forma experimental no afã de reproduzir comportamentos observados.

Como escreve Pamela McCorduck (2004, p.487), refletindo sobre tendências observadas então desde a década anterior: "No momento em que escrevo, a IA vive uma hegemonia Empertigada, pessoas que acreditam que inteligência de máquina, pelo menos, é melhor expressada em termos lógicos, e mesmo matemáticos".[8]

A década de 2000 testemunhou um avanço muito grande em IA, embora a sociedade não percebesse ainda o que estava acontecendo nos laboratórios (isso só acontece na temporada seguinte!). O trabalho em representação de conhecimento tornou-se cada vez mais teórico, tanto no seu estudo de ontologias e lógicas, quanto na exploração de métodos racionais de decisão e planejamento para um ou mais agentes. Acima de tudo, o que caracterizou essa década foi o crescimento constante das técnicas coletivamente designadas por "aprendizado de máquina".

A palavra "aprendizado" inclui muitos diferentes fenômenos; na maior parte do século passado, "aprendizado de máquina" se referia a um conjunto difuso de ideias cujo objetivo era melhorar o desempenho de um sistema com base em suas experiências. Por exemplo, um programa poderia se beneficiar de notas inseridas pelo usuário após sua execução. Ou poderia interagir com o usuá- rio, recebendo instruções e se adaptando de acordo. Ou poderia processar uma base de dados para decidir a melhor maneira de classificar imagens. Esse mosaico de conceitos era prevalente pelo fim do século XX (Mitchell, 1997). A partir daquele momento, houve um extraordinário crescimento na quantidade de informações e dados disponíveis publicamente, um incrível aumento no poder de coletar, transmitir e processar esses dados, e um substancial interesse em algoritmos capazes de usar dados para resolver problemas práticos. Subitamente, a pesquisa em "mineração de dados" se tornou extremamente valorizada tanto na academia quanto na indústria. Em um sentido bastante claro, o trabalho continuava com uma perspectiva empertigada, pois procuravam-se fundamentos matemáticos sólidos, por exemplo ligados a análises estatísticas. Ao mesmo tempo, aplicações de todas as naturezas passaram a fazer parte do universo do aprendizado de máquina, tornando a área muito diversa e abrangente.

Parece oportuno terminar esse exame de temporadas anteriores discutindo um artigo publicado em 2009 por três pesquisadores da empresa Google, Alon Halevy, Peter Norvig e Fernando Pereira, intitulado "The unreasonable effectivess of data" (Halevy et al., 2009). O argumento central do artigo é que muito se pode extrair de grandes massas de dados, mais até do que a intuição poderia sugerir. Assim, os melhores programas de tradução de textos usam estatísticas extraídas de grandes conjuntos de textos para mapear palavras de uma língua para outra – compreender de fato os textos não parece ser necessário se muitos dados são utilizados.[9] Dada a complexidade da linguagem, a procura por gramáticas e ontologias manualmente construídas parece ser menos eficaz que a extração de padrões linguísticos de grandes bases de texto. Em resumo: é incrível (*unreasonable!*) o quanto uma máquina pode parecer inteligente sem ter nenhuma real compreensão do que está fazendo, desde que o faça com base em padrões extraídos de muitos dados! Um modelo extremamente flexível (desgrenhado?) terá melhor desempenho do que um modelo dogmático (empertigado?) se puder se basear em um oceano de observações. E com essa surpreendente observação, passamos para a temporada atual.

A nova temporada

Uma nova temporada em IA começou por volta de 2010 e ainda não terminou. Para os propósitos deste artigo, uma grande parte do que ocorreu na última década pode ser resumido de forma breve. Grandes bases de conhecimento foram construídas, muitas das quais se beneficiando de representações baseadas em linguagens formais. A lógica clássica foi estendida em muitas direções, e esse processo passou a ser abordado de maneira formal e organizada. Métodos de manipulação de incertezas e de tomada de decisão também se consolidaram em torno de núcleos axiomatizados e formais. A teoria de aprendizado de máquina, em diálogo permanente com a disciplina de estatística, se tornou extremamente sofisticada. Em muitos sentidos, os empertigados tiveram grandes vitórias – não apenas em garantir que um núcleo duro da disciplina se aglutinasse de forma organizada, mas também em prover arquiteturas baseadas em princípios gerais de racionalidade.

Porém, a grande sensação dessa temporada foi o extraordinário desempenho obtido por métodos baseados em aprendizado de máquina, e em particular em aprendizado profundo (do inglês *deep learning*). Essa última expressão foi cunhada para se referir a modelos especificados por meio de um grande número de camadas, mais precisamente "redes neurais com muitas camadas". Embora vários resultados já tivessem sido publicados com tais redes antes de 2012, naquele ano a comunidade se surpreendeu quando uma particular rede neural profunda venceu uma competição de classificação de imagens (Krizhevsky et al., 2012). Nos anos seguintes redes neurais cada vez maiores e mais profundas continuaram surpreendendo a comunidade, atingindo desempenho melhor que o humano em várias tarefas envolvendo imagens e textos. Em particular, tarefas

como tradução automática passaram a ser dominadas por redes neurais (muito) profundas, em alguns casos com milhões ou bilhões de parâmetros estimados a partir de grandes massas de dados (Devlin et al., 2019; Tan; Quoc, 2019).

O resultado foi uma revolução em IA. Computadores passaram a ter desempenho melhor que o humano em tarefas como geração de respostas em *chatbots*, rotulagem de imagens em redes sociais, produção de temas musicais. Alguns desafios cuja solução parecia distante, como o controle de carros autônomos ou a capacidade de jogar Go, foram vencidos em um curto espaço de tempo. Tal sucesso não passou despercebido de engenheiros, médicos, advogados, economistas; em poucos anos todos os problemas que envolvem automação passaram a receber atenção via *deep learning*. A literatura nesse tópico cresceu imensamente, e muitos artigos rapidamente receberam centenas, em alguns casos milhares, de citações. Profissionais experientes em *deep learning* passaram a ser disputados no mercado internacional.

É interessante notar que redes neurais, após um áureo período na década de 1990, perderam boa parte de seu apelo aproximadamente entre 2000 e 2010. De forma simplificada, pode-se dizer que as redes neurais do ano 2000 foram atacadas por dois lados. Por um lado, as redes neurais eram pouco transparentes quando se tratava de representar conhecimento, e foram suplantadas nessa tarefa por modelos probabilísticos como redes Bayesianas. Por outro lado, as redes neurais dos anos 2000 não conseguiam competir, em termos de acurácia em atividades de classificação, com modelos estatísticos otimizados como Máquinas de Vetores de Suporte (Hastie et al., 2009). Apesar disso, alguns pesquisadores, acreditando que a melhor forma de reproduzir comportamento cerebral seria investir em redes de neurônios artificiais, continuaram a refinar algoritmos e cálculos, finalmente obtendo um grande sucesso. Pesquisadores como Geoffrey Hinton, Yann LeCun, Youshua Bengio, Jurgen Schmidhuber venceram ao manter uma abordagem resolutamente desgrenhada.[10]

Essa verdadeira mudança de paradigma é um ponto central neste artigo. Subitamente, IA se tornou um campo focado em modelos neurais desgrenhados. Note que existe uma questão metodológica complexa a respeito das recentes redes neurais. Cada vez mais essas redes neurais dependem de complicadas técnicas numéricas para que consigam processar montanhas de dados; qualquer semelhança que poderia ser pretendida com sistemas neuronais reais foi abandonada nessa jornada. Por exemplo, redes neurais designadas como *transformers,* que hoje obtêm alguns dos melhores resultados em IA (Devlin et al., 2019), atendem a uma variedade de requisitos e intuições sem relação direta com estruturas observadas no cérebro. Em resumo, trata-se de uma abordagem realmente desgrenhada, sem uma conexão biológica que muitas vezes existiu. Além disso, qualquer possibilidade de compreender o funcionamento dessas redes neurais gigantes a partir de princípios simples parece no momento descartada. A mencionada citação de Aaron Sloman (1990, p.2) é hoje mais verdadeira ainda: uma

rede neural "geralmente depende de uma rede de conexões totalmente opaca entre unidades de processamento e pesos que modificam suas influências entre si". A opacidade de redes neurais se tornou uma preocupação significativa nos últimos anos, por prejudicar a interpretabilidade esperada de sistemas ocupados com decisões práticas (Darpa, 2016). Angústias que haviam sido esquecidas na década de 1990, relativas à dificuldade de compreender redes neurais, retornaram com força. Além disso, a literatura tem apontado falhas de redes neurais profundas que ocorrem sem que seja possível entender sua razão (Marcus; Davis, 2019).

Seja como for, não há como negar que tarefas antes consideradas insolúveis foram vencidas com uso de redes neurais profundas construídas a partir de grandes bases de dados. Redes neurais profundas têm sido particularmente bem-sucedidas na operação chamada *end-to-end*: a rede recebe como entrada os dados crus e entrega na saída a decisão final. Por exemplo, a entrada da rede é uma imagem inteira e a saída indica se um ator famoso está na imagem ou não; isso é feito sem que nenhuma outra informação seja extraída da imagem (por exemplo, informação sobre cores, sobre luminosidade, sobre número de pontos pretos etc.). Outro exemplo: um texto é fornecido na entrada e na saída aparece um sumário do texto, sem que internamente seja construída nenhuma representação do conteúdo do texto de entrada. Hoje parece estar em curso uma corrida por redes neurais profundas que possam realizar mais e mais tarefas de forma *end-to-end*, incluindo por exemplo recursos de memória e recursão (Graves et al., 2014). Grande parte dos pesquisadores ligados a aprendizado profundo parece operar segundo a crença de que, com mais e mais dados, será possível aprender redes neurais que resolvam problemas de forma abrangente, onde toda entrada gera uma saída inteligentemente selecionada.[11] Isso será a vitória dos desgrenhados; caberá aos empertigados explicar essa vitória de forma organizada em livros-texto.

Assim, a IA vive uma encruzilhada que pode ser entendida pelo debate entre desgrenhados e empertigados.

Pode ser que a inteligência artificial seja desgrenhada não só por ser complexa (toda inteligência provavelmente será complexa), mas porque será obtida por um processo empírico onde uma inteligência "crescerá" a partir de dados. Ou pode ser que as dificuldades hoje enfrentadas por aprendizado profundo se tornem insustentavelmente pesadas. Pode ser que a necessidade cada vez maior de dados se torne inviável: faz sentido coletar um milhão de vídeos de pessoas fritando ovo para aprender a fritar um ovo – ou é melhor simplesmente pedir por instruções formais sobre como fritar um ovo? Pode também ser que haja de fato necessidade de um sistema físico de símbolos para que uma real inteligência seja obtida. E se for necessário construir uma rede neural que contenha no seu bojo um sistema de símbolos: seria essa a maneira mais eficiente de obter inteligência artificial?

Nos últimos anos tem havido um considerável esforço para unir aprendizado de máquina, e em particular aprendizado profundo, com técnicas "clássicas" de IA (ou seja, com técnicas empertigadas). Um consenso ainda não emergiu sobre como fazer essa união funcionar. Seria o caso de construir redes neurais que conseguem realizar operações lógicas? Procurar uma síntese neuro-simbólica? Ou desenvolver técnicas de aprendizado de máquina que consigam receber instruções em alto nível de abstração, como nós humanos recebemos na escola? A literatura está congestionada de propostas. Pode ser que uma delas faça a IA retornar para uma supremacia empertigada, mas não é claro que isso poderia acontecer.

Para encerrar esta seção, passo a oferecer algumas considerações em primeira pessoa. Primeiro, é natural que se busque ampliar o poder de representação de redes neurais (e também de outras técnicas usadas em aprendizado de máquina). Nada mais natural do que ampliá-las mediante a combinação com conceitos de linguagens formais e de análise matemática. É possível que redes neurais cada vez mais poderosas, e aumentadas com melhores técnicas de representação, consigam evoluir, e em algum ponto no futuro reproduzam de forma *end-to-end* todo o poder observado na inteligência humana. Essa será uma vitória que os desgrenhados celebrarão, e será uma vitória que nos ajudará a entender melhor o que é a inteligência e o que significa ser inteligente. Mas não creio que esse caminho, certamente interessante pelo que nos dirá sobre nós mesmos humanos, seja o mais eficiente em termos pragmáticos para a construção de inteligências artificias. A quantidade de dados e o poder computacional necessários serão imensos; seu custo será gigantesco. Além disso, a dificuldade de manipular objetos computacionais tão opacos como redes neurais profundas não podem ser desprezadas.

Em vista disso, penso ser mais promissor focar em sistemas inteligentes compostos de vários módulos, alguns baseados em dados e intrinsecamente desgrenhados, enquanto outros são baseados em princípios lógicos ou racionais claros. Os módulos desgrenhados serão avaliados com princípios estatísticos, verificando seus erros e acertos (como hoje testamos remédios ou vacinas). O desafio da projetista humano será combinar tais módulos em uma arquitetura coerente e eficiente.[12] Em sua história a IA já viu muitas arquiteturas amplas serem propostas; é hora de revisitá-las e combiná-las com os mais recentes e desgrenhados avanços.

Próximas temporadas?

No momento, ninguém espera que a IA tenha sua próxima temporada cancelada; pelo contrário, as expectativas da sociedade sobre essa tecnologia são imensas. Um pouco de excesso pode ser observado: em alguns casos excesso de propaganda para produtos ditos de IA e em outros casos excesso de preocupação com efeitos da tecnologia. A comunidade envolvida no desenvolvimento da IA, em sua grande maioria, segue procurando melhorar a produtividade e quali-

dade da vida humana mediante a construção de artefatos que possam nos auxiliar inteligentemente. Como atingir esse objetivo é a questão: estamos em busca de artefatos racionais baseados em princípios claros, ou artefatos empíricos que reproduzem padrões? Essa é, no fundo, uma das principais tensões hoje na área de IA. A sugestão aqui apresentada é que precisamos investir em arquiteturas baseadas em princípios de racionalidade e que permitam abrigar vários módulos simultaneamente, muitos dos quais baseados em coleta maciça de dados. Para conferir se essa sugestão de fato terá sucesso, só assistindo a próxima temporada.

Notas

1 No original: "*An excessive preoccupation with formalism is impeding the development of computer science*."

2 No original: "Generality in computer science". O conteúdo da palestra original não foi publicado; porém McCarthy (1987) publicou mais tarde um artigo comentando a palestra original.

3 No original: "*paper will discuss programs to manipulate in a suitable formal language (most likely a part of the predicate calculus) common instrumental statements*". Logo em seguida McCarthy indica que o projeto seria em conjunto com Marvin Minsky; porém na versão do artigo distribuída em 1996 (no site www-formal.stanford.edu/jmc/mcc59.pdf), McCarthy declara: "*This was wishful thinking. Minsky's approach to AI was quite different*".

4 McCarthy (1974) critica a doença do "*look ma, no hands*" em um comentário publicado em 1974, se referindo à situação em que alguém programa algo em um computador e publica um artigo anunciando que um computador obteve sucesso.

5 No original: "*to understand cognition by a formal analysis of language seems to me like trying to understand baseball by an analysis of the physics of what happens when an idealized bat strikes an idealized baseball*".

6 Citação completa no original: "*Here we have a branch of AI (yes, it is part of AI, not a new rival discipline), that is heavily mathematics-based, yet, although the general principles on which a particular network learns during its training period may be well understood, the operation of the final system when applied to real tasks generally depends on a totally opaque network of connections between processing units and weights that modify their influence on one another*".

7 Nesse caso, posso relatar em primeira pessoa, tendo feito um curso sobre IA baseado no livro de Rich e Knight na Carnegie Mellon University em 1992. As dificuldades com o livro-texto não pareciam ser falha dos autores; o problema estava no confuso estado da disciplina de IA naqueles dias.

8 No original: "*As I write, AI enjoys a Neat hegemony, people who believe that machine intelligence, at least, is best expressed in logical, even mathematical terms*". Pamela McCorduck escreve esse trecho em 2004 ao publicar a segunda edição do seu celebrado livro sobre a história da IA; a primeira edição do livro cobria fatos até 1977. O uso de *neat* e *scruffy* por McCorduck não é sempre claro e revela dificuldades dessa nomenclatura; por exemplo, o projeto CyC, que pretende montar uma grande base de sentenças lógicas codificando aspectos do mundo real, e portanto, em um objetivo

estruturante e baseado em princípios teóricos, é designado *scruffy* por ser tão grande que é praticamente capturar toda sua funcionalidade. De qualquer forma, a citação indicada acima está alinhada com os argumentos do presente artigo.

9 Considere o seguinte comentário de McCorduck (2004, p.223), relativo à tradução computacional nas décadas de 1960 e 1970: "*It was soon obvious that translation isn't merely transformation, but consists of a process of 'world modeling,' as Yehoshua Bar-Hillel, the well-known Israeli linguist, put it — the machine must, in some sense, understand the text before it can translate into another language, and it is in reference to the world model that understanding takes place*".

10 Considere o seguinte trecho na revista *Science*: "*Yann LeCun, Facebook's chief AI scientist in New York City, worries that shifting too much effort away from bleeding-edge techniques toward core understanding could slow innovation and discourage AI's real-world adoption. 'It's not alchemy, it's engineering,' he says. 'Engineering is messy*'" (Hudson, 2018).

11 Um artefato que atinge inteligência geral dessa forma claramente viola a famosa Hipótese do Sistema Físico de Símbolos cristalizada por Newell e Simon em 1975, quando ambos receberam o Turing Award: "The Physical Symbol System Hypothesis. *A physical symbol system has the necessary and sufficient means for general intelligent action*" (Newell; Simon, 1976, p.116). Newell e Simon (1976, p.116) esclarecem o uso do adjetivo *physical* como segue: "*The adjective 'physical' denotes two important features: (1) Such systems clearly obey the laws of physics – they are realizable by engineered systems made of engineered components; (2) although our use of the term 'symbol' prefigures our intended interpretation, it is not restricted to human symbol systems*".

12 Um foco renovado em arquiteturas (revisitando arquiteturas do passado) foi proposto por Scott Sanner em uma palestra disponível em <http://c4ai.inova.usp.br/contact/> (youtube).

Referências

ABELSON, R. P. Constraint, construal, and cognitive science. *Third Annual Conference of the Cognitive Science Society*, p.1-9, 1981.

BERGER, A. L. et al. The Candide system for machine translation. *Workshop on Human Language Technology*, p.157-62, 1994.

BROOKS, R. Elephants don't play chess. *Robotics and Autonomous Systems*, v.6, p.3-15, 1990.

CHEESEMAN. P. In defense of probability. *International Joint Conference on Artificial Intelligence*, p.1002-9, 1985.

DARPA. *Explainable Artificial Intelligence (XAI),* DARPA-BAA-16-53, 2016.

DAVIS, R. et al. What is a knowledge representation? *AI Magazine*, v.14, n.1, p.17-33, 1993.

DEVLIN, J. et al. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.* arXiV:1810.04805, 2019.

GRAVES, A. et al. *Neural Turing Machines*, arXiV:1410.5401, 2014.

HALEVY, A. et al. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, p.8-12, 2009.

HASTIE, T. et al. *The Elements of Statistical Learning*. Springer, 2009.

HAYES, P. In defence of logic. *International Joint Conference on Artificial Intelligence*, p.559-65, 1977.

HUDSON, M. IA researchers allege that machine learning is alchemy. *Science,* 2018.

KRIZHEVSKY, A. et al. ImageNet classification with deep convolutional neural networks. *Neural Information Processing Systems*, p.1106-14, 2012.

MARCUS, G.; DAVIS, E. *Rebooting AI – Building Artificial Intelligence We Can Trust*. Pantheon, 2019.

MCCARTHY, J. Programs with common sense. *Mechanisation of Thought Processes (Volume I)*, p.75-84, 1959.

_____. Review of "Artificial Intelligence"A General Survey". *Artificial Intelligence*, v.5, n.3, p.371-22, 1974.

_____. Generality in artificial intelligence. *Communications of the ACM*, p.257-67, 1987.

MCCORDUCK, P. *Machines Who Think*. s.l.: CRC Press, 2004.

MINSKY, M. Form and contente in computer science. *Journal of the Association for Computing Machinery*, v.17, n.2, p.197-215, 1970.

_____. *The Society of Mind*. s.l.: Simon & Schuster, 1986.

MITCHELL, T. *Machine Learning*. s. l.: McGraw Hill, 1997.

NEWELL, A. *Unified Theories of Cognition*. s. l.: Harvard University Press, 1990.

NEWELL, A.; SIMON, H. A. *Computer Science as Empirical Enquiry: Symbols and Search. Communications of the ACM,* v.19, n.3, p.113-26, 1976.

NILSSON, N. J. Artificial intelligence prepares for 2001. *AI Magazine*, v.4, p.7-14, 1983.

_____. *The Quest for Artificial Intelligence*. s. l.: Cambridge University Press, 2009.

PEARL, J. *Probabilistic Reasoning in Intelligent Systems*: Networks of Plausible Inference. s. l.: Morgan Kauffman, 1988.

RICH, E.; KNIGHT, K. *Artificial Intelligence*. s. l.: McGraw Hill, 1991.

_____. *Artificial Intelligence*. 3.ed. s. l.: McGraw Hill, 2010.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating error. *Nature*, v.323, n.6088, p.533-536, 1986.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence*: A Modern Approach. s. l.: Prentice Hall, 1995.

SERGIOVANNI, T. Mystics, neats and scruffies: Informing professional practice in educational administration. *Journal of Educational Administration*, v.27, n.2, p.7-21, 2007.

SLOMAN, A. Must inteligente systems be scruffy? *Evolving Knowledge in Natural Science and Artificial Intelligence*, Pitman, 1990.

TAN, M.; QUOC Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning,* p.6105-14, 2019.

WILKS, Y. *An Artificial Intelligence Approach to Machine Translation*. Tech report AD0741199, Stanford University, 1972.

RESUMO – O estudo de Inteligência Artificial (IA) tem sido perseguido, desde seu início, segundo dois estilos diferentes, jocosamente referidos como *scruffy* (desgrenhado) e *neat* (empertigado). Esses estilos na verdade refletem distintas visões sobre a disciplina e seus objetivos. Neste artigo revisamos a tensão entre desgrenhados e empertigados ao longo da história da IA. Analisamos o impacto do atual desempenho de métodos de aprendizado profundo nesse debate, sugerindo que o desenvolvimento de arquiteturas computacionais amplas é um caminho particularmente promissor para a IA.

PALAVRAS-CHAVE: Inteligência Artificial, Lógica, Representação de conhecimento, Aprendizado profundo.

ABSTRACT – The study of Artificial Intelligence (AI) has been pursued from the very beginning in two different styles, jokingly referred to as *scruffy* and *neat*. These styles actually reflect distinct viewpoints of the discipline and its objectives. In this paper, we review the tension between scruffies and neats over the history of AI. We analyze the impact of current deep learning methods in this debate, suggesting that the development of broad computational architectures is a particularly promising path for AI.

KEYWORDS: Artificial Intelligence, Logic, Knowledge representation, Deep learning.

Fabio G. Cozman é professor titular da Escola Politécnica da Universidade de São Paulo (USP) e diretor do Centro de Inteligência Artificial (C4AI) na USP. Foi coordenador do Comitê Especial em Inteligência Artificial da Sociedade Brasileira de Computação. @ – fgcozman@usp.br / https://orcid.org/0000-0003-4077-4935.
Recebido em 10.3.2021 e aceito em 11.3.2021.

[I] Universidade de São Paulo, Escola Politécnica, São Paulo, Brasil.