O problema da explicação em Inteligência Artificial:

considerações a partir da semiótica

Joel Carbonera[1]
Bernardo Gonçalves[2]
Clarisse de Souza[3]

Resumo: Desde os sistemas especialistas dos anos 1980 e 1990, pesquisadores de Inteligência Artificial (IA) dedicam-se ao problema da explicação, a saber, dada uma inferência por parte do sistema, como identificar os passos ou mecanismos que o levaram a tal conclusão. Com o recente sucesso dos sistemas de IA atuais, sobretudo os baseados em aprendizagem profunda, esse problema voltou à tona com vigor, agora mais pronunciado, por eles serem opacos quanto ao seu processo de inferência, em contraste com os sistemas especialistas, então baseados em regras lógicas. Neste texto, apresentamos o problema da explicação, incluindo destaques de sua literatura mais recente na área de IA. Em seguida, indicamos lacunas de abordagens passadas e recentes, e apresentamos então considerações a partir da semiótica de Peirce que, conforme argumentamos, poderiam contribuir para uma condução equilibrada dessa tecnologia na sociedade.

Palavras-chave: Inteligência Artificial. Explicabilidade. Semiótica e Pragmatismo. Engenharia Semiótica.

Abstract: Since the expert systems of the 1980s and 1990s, Artificial Intelligence (AI) researchers have tried to solve the the problem of explanation, namely, given an inference from the system, how to identify the steps or mechanisms that have led to the conclusion. With the recent success of AI systems, especially those based on *deep learning*, this problem has come to the fore again more forcefully since the processes are opaque as far as their inferences are concerned, in contrast to expert systems, which are based on logical rules. In this text, we present the problem of explanation, including highlights from its most recent literature in the area of AI. Next, we indicate gaps in past and recent approaches, and then present considerations from Peirce's

[1] Doutor em Ciência da Computação na Universidade Federal do Rio Grande do Sul, membro do grupo BDI (grupo de bancos de dados inteligentes) da UFRGS e do grupo de trabalho financiado pelo IEEE RAS, intitulado Padrão para Ontologias para Robótica e Automação (IEEE RAS WG ORA), coordenador de padronização do campo de Robótica e Automação no capítulo IEEE South Brazil Robotics & Automation Society. E-mail: jlcarbonera@inf.ufrgs.br.
[2] Pós-doutorado na Universidade de Michigan–Ann Arbor, Ph.D. em Modelagem Computacional com foco em Data Science/ Laboratório Nacional de Computação Científica (LNCC), doutorando em Filosofia da Ciência/ USP, membro da Associação Profissional de Scientiae Studia e da Associação para a Filosofia e História da Ciência do Cone Sul. E-mail: bgoncalves1@gmail.com.
[3] Professora titular do Departamento de Informática PUC-Rio, doutora em Linguística Aplicada (foco interação humano - computador), Criadora da Engenharia Semiótica, em 2010 foi agraciada com o ACM SIGDOC Rigo Award e em 2013 tornou-se membro da ACM SIGCHI CHI Academy. Em 2014 recebeu o título de HCI Pioneer, outorgado pelo Comitê Técnico de Interação Humano-Computador (TC13) da IFIP. Também em 2014 foi selecionada como uma das 52 pesquisadoras mulheres a figurarem na primeira edição do CRA-W / Anita Borg Institute Notable Women in Computing Card Deck. Em 2016 recebeu o Prêmio do Mérito Científico da SBC e em 2017 o prêmio | Carreira de Destaque em IHC, concedido pela Comissão Especial de Interação Humano Computador da SBC. Em licença sabática da PUC-Rio, trabalhando como Pesquisadora Senior na IBM Research Brazil. E-mail: clarisse@inf.puc-rio.br.

semiotics, which, as we argue, could contribute to a balanced management of this technology in society.

Keywords: Artificial intelligence. Explanability. Semiotics and Pragmatism. Semiotic Engineering.

Introdução

Em dezembro de 2016, a jornalista Carole Cadwalladr (2016), do *The Guardian*, foi a um popular motor de busca para uma pesquisa, e digitou "j-u-d-e-u-s", seguido de "s-ã-o". As sugestões de "autocompletar" e os resultados obtidos (o conjunto de páginas retornadas) foram surpreendentes. Então, numa nova busca, ela digitou "m-u-ç-u-l-m-a-n-o-s", e novamente "s-ã-o"; noutra, digitou "m-u-l-h-e-r-e-s", depois "s-ã-o"; e assim por diante, até se defrontar com um mundo onde "Hitler foi um cara legal" (sic!). Ocorre que o sistema de busca em questão pode ser considerado um sistema de Inteligência Artificial (IA) que *aprende* dos dados, e que é potencialmente vulnerável a vieses dos mais inofensivos aos mais repulsivos. Mas como saber que tipo de viés pode estar codificado em um sistema de IA? Essa é uma questão que está radicada no chamado problema da explicação de (um sistema de) IA – a saber, *como identificar os passos ou mecanismos que levaram um sistema a chegar a tal ou qual decisão?* –, que é o tema deste texto.

À medida que vão sendo implantados em nossa rotina uma miríade de sistemas de IA, das recomendações de produtos, ao reconhecimento facial e aos *chatbots*, a sociedade tem despertado para o problema da lacuna de explicações acerca do comportamento (inteligente) de sistemas de IA. São sinais disso: a formação, em novembro de 2016, de um consórcio de parceria em IA para beneficiar as pessoas e a sociedade (HERN, 2016), por parte de algumas das maiores empresas de tecnologia do mundo (Google, Facebook, Amazon, IBM, Microsoft); e a legislação relativa à *General Data Protection Regulation* (GDPR), que entrou em vigor em 2018 no âmbito da União Européia, induzindo um direito à explicação por parte da/do cidadã(o) que seja "afetado significativamente" (sic!) por decisões automatizadas tomadas por algoritmos preditivos no nível (individual) de um usuário (GOODMAN, 2016).

No contexto dessa recente guinada com relação à relevância e à seriedade com que o tema da IA é tratado no domínio público, a partir do ano de 2017, a comunidade

técnica de IA e aprendizagem de máquina tem reagido com a criação de fóruns especializados (de realização conjunta com as conferências ou simpósios técnicos regulares) para discussão da questão da explicação em IA, o que deu origem à expressão *Explainable AI*, também conhecida pelo acrônimo XAI. Como veremos neste texto, entretanto, é possível que tais iniciativas ainda se ressintam de uma perspectiva demasiado unilateral (oriunda das virtudes e dos vícios de uma orientação técnica específica da comunidade científica), havendo espaço então – sobretudo, em se tratando de um problema de explicação, ou, se assim o quisermos, de comunicação de sentido – para que seja ampliada a discussão à luz das ciências humanas e da semiótica.

Começaremos com uma breve apresentação do problema da explicação em IA e aprendizagem de máquina, seguida de uma breve revisão da literatura pregressa e recente no tema. Procederemos então, indicando lacunas na maneira como o problema vem sendo abordado em IA, e apresentaremos uma perspectiva mais ampla do problema a partir da semiótica de Peirce, em direção a uma condução equilibrada da implantação de sistemas de IA na sociedade.

Visão geral do problema da explicação em IA

Uma breve retomada histórica

Entre as décadas de 70 e 80, a capacidade de explicação de inferências se apresentou como um problema relevante que atraiu a atenção de pesquisadores de IA. Isso se deu primeiro no contexto dos chamados "sistemas especialistas", que eram baseados em regras lógicas e heurísticas de busca para chegarem a uma conclusão visando apoiar a decisão de especialistas humanos. Esse é o caso, por exemplo, do apoio ao diagnóstico médico (CLANCEY; SHORTLIFFE, 1984). Para que a conclusão do sistema fosse aceita, era preciso oferecer ao ser humano responsável uma espécie de rastro do raciocínio automático, identificando os passos tomados pela dedução lógica e os fatos por ela empregados – por exemplo, apresentar a regra de que a hemodinâmica é aceitável *se* a frequência cardíaca é aceitável, a frequência do pulso é estável o suficiente, e a pressão sanguínea sistólica é aceitável etc. (ibid., p. 246).

Em 1985 surgiram as redes bayesianas, que são modelos gráficos probabilísticos (PEARL, 1988). Nelas, eventos são associados a uma probabilidade e conectados com direcionalidade a outros eventos. Por exemplo, seja "chuva" um evento com probabilidade *p1* e "grama molhada" um evento com probabilidade *p2* condicionada por *p1*. Essa proposição pode ser representada visualmente como um grafo direcionado que leva de "chuva" a "grama molhada". Isso, aliado ao fato dos eventos possuírem um nome inteligível (como já era o caso dos sistemas especialistas), pode ter contribuído para o problema de a explicação não ter assumido maior relevância na ocasião.

Entre 1990 e 2000, então com o surgimento dos sistemas de recomendação, novos tipos de inferência necessitavam de explicação. Um cenário típico é a recomendação de um filme a um usuário porque o filme foi bem avaliado por um outro usuário que vem a ser "amigo" do primeiro. Estudos como o de Herlocker et al. (2000), contemplando variados formatos de explicação para esse tipo de inferência, confirmaram que os usuários achavam de fato necessária a apresentação de uma explicação. Eles indicaram também que formatos mais simples e conclusivos de explicação – por exemplo, mostrar a nota (digamos, 4 de 5 estrelas) dada pelo outro usuário (amigo), bem como indicar uma propriedade marcante do filme, como a presença de um ator favorito – eram preferíveis a formatos de explicação baseados em conceitos de aprendizagem de máquina – como a estimativa de confiança do modelo preditivo. Esses estudos (cf. levantamento feito por BIRAN; COTTON, 2017) são informativos, pois sugerem direções acerca do tipo de explicação capaz de satisfazer um (a) usuário (a).

Explicação de aprendizagem profunda: o elemento novo

A expressão "aprendizagem profunda" foi cunhada na comunidade de aprendizagem de máquina por Dechter em (1986) e empregada pela primeira vez na comunidade de redes neurais artificiais por Aizenberg e outros no ano 2000. Em seguida ela se tornou especialmente popular no contexto das redes neurais profundas (DNNs, do inglês *deep neural networks*), que são talvez os modelos mais bem-sucedidos de aprendizagem de máquina até hoje. Apesar das DNNs já existirem há mais tempo

como parte de uma ampla classe de modelos considerados do tipo caixa preta, na última década houve uma retomada de interesse por tais abordagens. Isso foi motivado principalmente pelo aumento do poder computacional disponível e pela disponibilização de um grande conjunto de dados devidamente classificados por seres humanos, tais como a ImageNet (DENG et al., 2009). Desde então, DNNs têm sido aplicadas com sucesso em diversos cenários de uso, obtendo resultados comparáveis aos obtidos por seres humanos em tarefas como o reconhecimento de objetos em imagens (HE et al., 2015). Isso trouxe à tona novamente, talvez com mais vigor, o problema da explicação, visto que, em aplicações reais, seres humanos (incluindo projetistas, usuários etc.) desejam saber como e/ou por que certos resultados foram obtidos.

As DNNs são redes neurais que possuem múltiplas camadas intermediárias de neurônios – conectados de camada a camada, com um valor numérico associado como peso que é ajustado via treinamento para modular a propagação do sinal recebido –, e podem ser treinadas para realizar uma tarefa computacional (por exemplo, classificação de imagens de animais). Para viabilizar o processo de treinamento, é necessário um conjunto suficientemente grande de dados, no qual cada entrada (imagem) do conjunto deve estar previamente associada a um rótulo (por exemplo, “gato”), que representa a resposta que se esperaria da rede neural para a dada entrada. Esse processo de treinamento visa ajustar os pesos das conexões entre os neurônios da rede de tal forma que ela seja capaz de mapear uma nova entrada, para a qual não se conhece o rótulo, a uma resposta correta. Ou seja, considerando este exemplo, o conhecimento sobre o padrão que representa cada animal fica implicitamente representado no conjunto de pesos da rede neural, sem a necessidade de se informar explicitamente ao sistema a lógica e os conceitos subjacentes a este conhecimento.

Do ponto de vista do problema da explicação, considerando a tarefa de classificar uma imagem como contendo (sim ou não) uma ave, podemos ilustrar o contraste entre um sistema de IA baseado em regras lógicas (cf. citado em subtítulo anterior deste artigo) com um baseado em redes neurais (digamos, DNNs). No primeiro caso, o sistema de IA poderia explicar que classifica o animal como ave porque ele tem

penas, asas, bico, duas patas etc. A inferência lógica é de que, se ele tem todos esses atributos, então se trata de uma ave. No segundo caso, a classificação de um animal como ave é função da similaridade que suas características (traços distintivos, até de caráter geométrico) têm com as características extraídas (automaticamente, por operações de processamento de imagem) de um vasto conjunto de imagens de aves. O sistema que "aprendeu" os traços distintivos das aves por meio de exemplos tentará identificar essas mesmas características nas novas imagens que for solicitado a classificar. Há dois pontos centrais no caso das DNNs: (i) as características aprendidas não possuem necessariamente uma relação perceptiva compatível com algo que um ser humano seria capaz de discernir (nomear) nas aves; e (ii) essa aprendizagem é autônoma, sem que seja necessário o acompanhamento de um ser humano. Ou seja, temos aqui dois aspectos convenientes do ponto de vista tecnológico, que se traduzem em um desafio peculiar do ponto de vista do problema da explicação e dos impactos da IA na sociedade.

Iniciativas recentes de pesquisa no problema da explicação em IA

A retomada de interesse no problema da explicação motivou o surgimento de diversos fóruns especializados dentro da comunidade de pesquisa de IA. Começamos pelos esforços conceituais e de levantamento bibliográfico, dedicados a mapear o problema e estruturá-lo por meio de distinções.

Biran e Cotton (2017), por exemplo, assumem que explicabilidade é algo fortemente relacionado à noção de interpretabilidade: um sistema interpretável seria aquele cujas *operações* são compreensíveis para nós humanos, seja por meio da inspeção do sistema, seja por meio de alguma explicação produzida durante o seu funcionamento. Eles estabelecem uma distinção (ibid.) entre interpretabilidade e a noção de justificação, cujo objetivo seria explicar por que a decisão tomada pelo sistema pode ser aceita como uma boa decisão. Ou seja, justificabilidade e interpretabilidade seriam capacidades complementares. Doran et al. (2017), por sua vez, identifica três classes de sistema: opacos, intepretáveis e compreensíveis. Sistemas opacos são como caixas pretas, i.e., seus mecanismos não são inspecionáveis por usuários. Sistemas interpretáveis, por outro lado, permitiriam inspeção, estudo e

compreensão dos seus processos e mecanismos internos, mesmo que essas tarefas demandem certo conhecimento técnico especializado. Já os sistemas compreensíveis seriam aqueles que, além de oferecerem um resultado, também oferecem símbolos que são inteligíveis para os usuários, permitindo compreender por que uma certa saída está associada a uma certa entrada. Segundo Doran et al. (ibid.), portanto, compreensibilidade e interpretabilidade seriam capacidades complementares. Lipton (2016), finalmente, afirma que outros autores negligenciam que explicabilidade não é uma noção absoluta, mas contextual. Com tal perspectiva, ele busca identificar propriedades desejáveis para sistemas interpretáveis, com destaque para *transparência*, relacionada à inteligibilidade dos mecanismos internos do sistema; e *interpretabilidade post-hoc*, relacionada à capacidade do sistema de oferecer informações úteis sobre seus resultados, para usuários diversos.

A maior parte das iniciativas discutidas nos fóruns especializados, entretanto, propõe abordagens (métodos, técnicas ou ferramentas) computacionais específicas, para lidar com o problema da explicabilidade. Para um exemplo do tipo de iniciativa focada naquilo que Biran e Cotton (2017) e Doran et al. (2017) chamaram de compreensibilidade, Yosinski et al. (2015) propõem duas abordagens para auxiliar (através de visualização) na compreensão dos processos internos realizados por modelos de redes neurais convolucionais (um tipo de DNN adequado para classificar imagens). A primeira abordagem (Fig. 1, Parte A) gera uma imagem que é capaz de identificar e realçar quais pixels de uma dada imagem de entrada geram os maiores níveis de ativação para um dado neurônio que o usuário deseja inspecionar. A segunda abordagem (Fig. 1, Parte B) permite gerar imagens sintéticas semelhantes a imagens naturais que representam quais são os padrões visuais que um certo neurônio aprendeu a detectar. Ambas as abordagens oferecem *insights* a respeito do funcionamento interno da rede neural.

Já Dhurandhar et al. (2018) propõem uma abordagem cujo objetivo é oferecer uma justificação (BIRAN; COTTON, 2017) para o resultado obtido por uma rede neural (sem necessariamente oferecer compreensibilidade a seus mecanismos ou processos internos). Considerando o problema da classificação de imagens de acordo com a classe de objetos que ela representa, por exemplo, segundo a perspectiva desses

autores, a justificação é elaborada em termos de dois tipos de características: (a) características que ela possui (evidências positivas), que seriam típicas da classe em que ela foi classificada e que seriam suficientes para justificar a classificação; e (b) características que ela não possui (evidências negativas), que seriam típicas de uma classe muito semelhante à classe em que ela foi classificada e cuja ausência seja necessária para justificar a classificação obtida. Por exemplo, conforme pode ser visto na Fig. 1 (Parte C), na tarefa de classificação de imagens de algarismos numéricos, seja uma imagem representando o algarismo quatro (e classificada por uma DNN de modo correspondente), a abordagem seria capaz de produzir uma visualização da imagem original que realçaria, com uma certa cor, pixels que suportam a identificação do dígito como "4" e, com uma outra cor, pixels ausentes na imagem, mas que se presentes, levariam a rede neural a identificar, digamos, o algarismo "9". A abordagem identifica ambos os conjuntos de pixels por meio da resolução de um problema de otimização. É importante notar que esta abordagem pode ser acoplada como um componente externo a uma rede neural, cuja finalidade é oferecer justificação para os resultados obtidos por ela.

É importante ressaltar, conforme diagnosticado por Biran (2017), que a maior parte dessas abordagens computacionais para o problema da explicação assume (em geral tacitamente) alguma definição do problema, não aprofunda discussões conceituais acerca do que viria a ser uma explicação, e não consideram com alguma profundidade aspectos sociais envolvidos. Há uma lacuna de estudos que visam identificar quais são os significados (e para quem significam, entre projetistas, usuários etc.) que devem ser preservados ao longo das cadeias de computação realizadas pelos sistemas de IA, e como viabilizar essa preservação.

Na próxima seção, articularemos essas lacunas à luz de conceitos semióticos introduzidos por Peirce (1955).

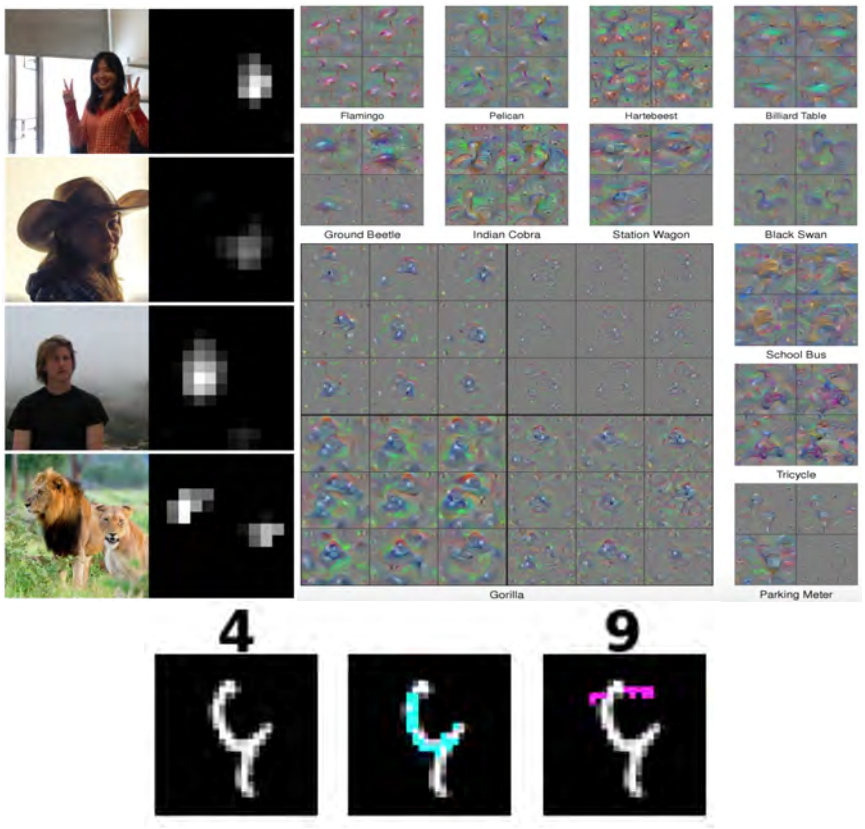

Figura 1. Parte A (superior, à esquerda): Representação do conjunto de pixels que gera um grande nível de ativação de um dado neurônio, para uma dada imagem de entrada. Quanto mais claro o pixel, maior é a ativação gerada no neurônio. Neste caso, é possível notar que o neurônio possui grandes níveis de ativação para faces, tanto humanas quanto de animais. **Fonte:** Yosinski et al. (2015). Reproduzido inalterado sob licença CC BY-NC-SA 3.0. **Parte B** (superior, à direita): Visualizações de imagens sintéticas geradas para representar o padrã o que cada um dos 12 neurônios da camada de saída (cada qual representando uma classe de objetos) detecta. Ao centro, são exibidas 4 visualizações para a classe Gorila, geradas por diferentes parametrizações do método, cada qual exibindo 9 imagens interpretáveis. Para as demais classes, são exibidas apenas 4 imagens interpretáveis selecionadas pelos autores. **Fonte:** Yosinski et al. (2015). Reproduzido inalterado sob licença CC BY-NC-SA 3.0. **Parte C** (inferior): À esquerda, vemos a imagem de entrada, juntamente com a classificação correspondente (algarismo 4) obtida por uma DNN. Ao centro, a visualização destaca em azul os pixels que fornecem evidência positiva à favor desta classificação. À direita, a visualização destaca em rosa os pixels que representam evidências negati vas contra uma classificação alternativa, mas visualmente muito próxima (algarismo 9). **Fonte:** Dhurandhar et al. (2018). Reproduzido com permissão do autor.

Lacunas da IA atual e da explicabilidade de seus sistemas

Uma peça de ficção escrita por Umberto Eco (1988) resume, com seu enredo alegórico, algo que parece estar fora do foco das iniciativas passadas e recentes de explicação em IA (e aprendizagem de máquina). O autor narra as peripécias de duas expedições interplanetárias de habitantes da Terra em um planeta fictício. Um grupo de expedicionários terráqueos toma e bebe – como se fosse água – um líquido extraterrestre que os habitantes do planeta chamaram de "água" na sua presença. Em

seguida, eles são vitimados por uma disenteria, desencadeando uma perigosa contenda diplomática. Para lidar com ela, uma segunda expedição de terráqueos foi enviada para examinar o comportamento dos nativos do planeta, buscando saber se o que eles chamavam de "água" era ou não era o que nós, terráqueos, assim chamamos.

A alegoria acima traz à tona, a distinção entre um símbolo (a palavra em si) e seu significado (seu sentido pragmático em contexto de uso). A alegoria oferece um cenário conveniente para ampliarmos a perspectiva acerca do problema da explicação em IA buscando estabelecer o limite das visões de explicação que revisitamos antes neste artigo.

Na continuação da narrativa de Eco (1988, p. 41), fica clara a dificuldade de comunicação entre os expedicionários humanos e os alienígenas, que costumam falar (e oferecer explicações) apenas em termos de seus estados neurais. Por exemplo, ao ver uma criança perto de um fogão quente, sua mãe extraterrestre assustada diria: "Ó meu Deus, ela vai estimular suas fibras-C!". Em outro exemplo, referindo-se ao que nós poderíamos caracterizar (ou "explicar") em termos como estes: "Aquilo pareceu um elefante, mas isso seria espantoso porque não se tem notícia de elefantes nesse continente; então me dei conta de que tem de ser um mastodonte" (ECO, 1988, p. 41).

Um nativo daquele planeta diria algo como: "Eu tive G-412, mas junto com F-11; e então eu tive S-147".

O último exemplo ilustra, ainda que de forma aproximada, o que está em jogo nas abordagens de explicação com base em regras lógicas (cf. anteriormente). É apresentada uma estrutura lógica instanciada por signos cujo sentido efetivo subjaz noutro nível semântico, correspondente a uma ontologia ou rede semântica específica e construída em certo contexto e (no caso de sistemas de IA) com um certo propósito. G-412 é o símbolo de uma noção conceitual de elefante, embora só possamos entender isto se articularmos dois (ou até mais) sistemas de significação. O mesmo acontece com a relação entre F-11 e a surpresa da constatação de que o lugar onde a percepção se dá é incompatível com alguma característica conceitual de G-412, e assim por diante. Para um computador, lembremos, nem G-412, nem "elefante", são *significativos* do ponto de vista conceitual. São somente cadeias simbólicas pertencentes a um vocabulário pré-definido sobre o qual se pode fazer alguma operação computacional programada.

Se consideramos as abordagens de explicação recentes voltadas para modelos caixa preta de aprendizagem de máquina, esse cenário fica ainda mais extremo. Em termos do exemplo de Eco, a assertiva análoga, nesse caso – desprovida de sua estrutura lógica dedutiva, e restrita a ocorrências correlacionadas –, seria o equivalente de "G-412, com F-11, com S-147".

Vemos então que, através da ficção, Eco (1988) sintetiza a essência de um debate que, apesar de muito antigo para a filosofia e estudos da linguagem ou da mente, volta a ser novamente central para a IA. Se por um lado, essa é uma discussão que tangencia questões das mais profundas, como a relação mente-corpo e o cartesianismo, ambas além do nosso escopo neste artigo, por outro lado, ela nos permite ressaltar – à luz da semiótica de Charles S. Peirce (1955) – dois elementos centrais disparados pela presença de um signo (re)conhecido: (i) o chamado processo de *semiose*, isto é, a construção de sentido entendida como indo além da fala ou da estrutura formal da linguagem, levando em conta aspectos pragmáticos (contextuais e de uso); e (ii) o raciocínio dito *abdutivo*, isto é, a inferência cuja justificação não se completa pela estrutura do argumento em si, como nos casos da indução e da dedução (na dimensão epistemológica), mas depende também da dimensão metodológica – seu papel e sua promessa para o avanço da investigação; por exemplo, se uma hipótese formada, digamos, por indução, é ou não é testável.

No caso dos expedicionários terráqueos que bebem o líquido extraterrestre porque os nativos o chamam de "água" (ibid., p. 41), é evidente a lacuna característica da semiose e do raciocínio abdutivo que se autocorrigem (SANTAELLA, 2004). A precipitação para uma conclusão na presença de evidências contingentemente consideradas suficientemente fortes, e ausência de contraditório, fez os expedicionários beberem "água" *daquele* planeta e terem disenteria. Agora, consideremos os sistemas de IA atuais, cujos modelos de comportamento são reconhecidamente desprovidos de correspondências claras com conceitos e regras lógicas que normalmente utilizamos numa explicação, mas têm entrado em contato com situações reais não antecipados – por exemplo, o caso mostrado pela jornalista (CADWALLDR, 2016). Não é difícil ver que a implantação desses sistemas em larga escala na sociedade, acompanhada das noções de explicação vista anteriormente neste

artigo, pode estar fadada a produzir um sem-número de frustrações, tais como as dos primeiros expedicionários da alegoria de Eco (1988).

Considerações sobre o potencial da semiótica de Peirce para a IA

Desde os sistemas de IA surgidos na última terça parte do século XX, a necessidade de explicação sempre foi evidente. O aspecto mais frequentemente mencionado é que a confiança nas decisões e avaliações produzidas por esses sistemas depende, como é frequente entre nós, seres humanos, de que o usuário – beneficiário ou de outra forma afetado pelo resultado da inferência automática – se *convença* de que ela procede e tem fundamento. Ora, mas de que depende convencermos alguém de que algo (seja o que for) procede e tem fundamento? Em geral, para qualquer agente (humano ou não) convencer uma pessoa de algo, é preciso que seja estabelecida uma relação entre as duas partes e que, em virtude dessa relação, sejam geradas e operacionalizadas (num processo de colaboração recíproca) inferências e expectativas que balizam um processo de comunicação. Idealmente, isso levará a uma persuasão e, como esperado, à confiança da pessoa no agente artificial. Ou seja, para se chegar propriamente à noção de explicação, há de se articular conceitos como o de relação (reciprocidade), comunicação e colaboração, em que inferências e expectativas desempenham um papel fundamental. É razoável considerar que estamos diante de questões pragmáticas muito claras.

Essa visão oferece um fio condutor e organizador de um espaço de pensamento que nos permite identificar, na prática de pesquisa ou de desenvolvimento de IA, determinadas rupturas que, a despeito da evolução das técnicas de raciocínio e aprendizado automáticos, não parecem ter sido vencidas. Uma questão central é a insistência em se manter *uma teoria do uso da IA divorciada de uma teoria da construção da IA*. Ora, os sistemas especialistas, dos anos 1980 e 1990, assim como os atuais sistemas autônomos que utilizam técnicas de aprendizagem profunda têm por destinação comum serem usados e serem úteis. A computação não é uma arte parnasiana, mas, sim, a base para processos de Engenharia de Sistemas e Tecnologias com que usuários humanos vão interagir por múltiplas razões e motivos, em múltiplas circunstâncias e, cada vez mais, em múltiplas modalidades e múltiplos dispositivos. A

rigor, a construção de sistemas de IA seria, por princípio, o estrito equivalente da construção de sistemas explicáveis, ou que podem ser explicados. No entanto, as divisões de competências, interesses, formações e práticas profissionais tanto na área científica quanto na grande área de desenvolvimento tecnológico reafirma o hiato que historicamente separou o estudo do *uso* de sistemas linguísticos (naturais ou artificiais) do estudo de sua estrutura ou sua *lógica*.

É notório que as novas técnicas de aprendizagem de máquina apresentam desafios ainda mais difíceis de explicabilidade e interpretabilidade do que enfrentaram os sistemas especialistas de duas ou três décadas passadas. Hoje, no entanto, ainda nos ressentimos da lacuna de uma infraestrutura teórica para unificar a modelagem, a engenharia e o uso de sistemas de IA. A semiótica peirceana permanece, no cenário contemporâneo, promissora. Recentemente, Nadin (2017) reiterou uma posição já anteriormente expressa em diversos textos, a saber, de que a IA não chegou ainda à realização e ao advento da "inteligência" artificial, embora tenha atingido resultados impressionantes na automação de tarefas associadas ao comportamento de seres naturais inteligentes. O ponto central do autor é a distinção entre *prever* (no âmbito do raciocínio sobre probabilidades e consequências lógicas) e *antecipar* (no âmbito dos significados possíveis, para agentes diversos). A seu ver, enquanto a computação supervaloriza a capacidade de prever o que vai acontecer, ela negligencia a capacidade de antecipar o que pode acontecer.

Entendemos que a questão do significado, ou da antecipação de oportunidades para além do raciocínio sobre probabilidades e consequências lógicas, nosso trabalho reunindo Semiótica e Computação pretende seguir uma outra linha, aderente à *Engenharia Semiótica* (DE SOUZA, 2005). A Engenharia Semiótica, em brevíssimas linhas, caracteriza a construção de sistemas computacionais como um processo de construção (Engenharia) Semiótica, centrado no conceito de *metacomunicação*. Partindo da premissa de que o destino de toda a computação é ser usada por pessoas em contextos pessoal e socialmente relevantes, os sistemas e tecnologias computacionais têm nas suas interfaces de usuário, o principal signo do que são. São elas que, por meio do desdobramento das interações que elas próprias facultam aos usuários, comunicam a esses usuários as formas, meios, efeitos, razões e possibilidades

de comunicação que os criadores (designers e engenheiros) de sistemas e tecnologias, eles sim, *anteciparam* ser relevantes, úteis, desejáveis, prazerosas e interessantes para os destinatários de seu trabalho. Sistemas e tecnologias, eles mesmos, conforme a visão de Nadin (2017), não fazem "antecipações de significado", apenas calculam os efeitos da presença ou ausência de padrões informacionais, seguindo instruções que direta ou indiretamente são estabelecidas por pessoas. Assim, a Engenharia Semiótica tem por objetivo restaurar o elo semiótico no processo de desenvolvimento de software, teorizando sobre a natureza metacomunicativa dos signos computacionais.

Para a IA contemporânea, a Engenharia Semiótica propõe, como um primeiro passo exploratório, uma investigação deste elo, buscando o rastro da significação humana sob as muitas camadas de sedimento computacional e informacional. Com um conjunto de métodos utilizados para a pesquisa sobre HCI (*Human–computer interaction*) e HCC (*Human-centered computing*) (SEBE, 2010), acreditamos ser possível sondar as antecipações de muitos seres humanos ligados à produção e ao consumo de informações (dados), programas (algoritmos) e sistemas (aplicações e tecnologias) usados em IA e, com isto, elaborar um enquadramento pragmático mais rico e consistente para o estudo de explicações e interpretações do comportamento de agentes artificiais inteligentes – seja para quem os utiliza (ou é afetado por ações e decisões de quem os utiliza), seja para quem os constrói (não importa em que ponto da potencialmente longa cadeia de produção de software que culmina nas tecnologias que hoje temos à nossa disposição).

Nosso trabalho, porém, é apenas um pequeno nicho de possibilidades que a semiótica, em particular a peircena, pode abrir para a IA. Voltando a um ponto anterior sobre a unificação teórica da pragmática com a lógica da descoberta e da antecipação, acrescentando agora os termos de Nadin (2017), a semiótica tem a possibilidade de apresentar-se como uma teoria integradora de que a computação aparentemente carece quando confrontada com a necessidade de produzir artefatos explicáveis e interpretáveis. Contudo, coloca-se em grande evidência e de imediato um grande desafio interdisciplinar. Para muitos pesquisadores, em muitas áreas de conhecimento e especialização que podem beneficiar-se significativamente de uma troca interdisciplinar, o discurso da semiótica como disciplina não é ameno, e por vezes não é

sequer compreensível. A efetiva fertilização do território da IA com as ideias de Peirce poderia ser levada adiante em larga escala como uma iniciativa interdisciplinar sustentada e sustentável que construa um território novo de interlocução científica em torno de projetos definidos e igualmente estimulantes para todas as disciplinas e subdisciplinas envolvidas.

De fato, a recente regulamentação europeia sobre o uso de dados e o direito à explicação (GOODMAN, 2016) tem o potencial de fomentar projetos interdisciplinares. Trata-se de um fato com repercussões sociais, legais e jurídicas, para não mencionar as econômicas e tecnológicas, que ainda mal vislumbramos. No entanto, já está claro que esta regulamentação, se não for ela mesma a causadora de grandes modificações no enquadramento social da IA – ou seja, do compromisso da computação com seus contextos de uso – pode ser o estopim de um movimento questionador profundo sobre os aspectos pragmáticos, éticos e filosóficos de teorias que até aqui se construíram sem pensar muito neles. Vemos, portanto, aí uma porta aberta para a circulação de ideias entre os domínios da computação e da semiótica, ancoradas a situações significativas e urgentes da sociedade contemporânea, cujas práticas já não conhecem uma fronteira clara entre o físico e o virtual, que, no entanto, se unificam sob o "signo" semiótico.

Enviado: 1 maio 2018
Aprovado: 29 maio 2018

Referências

AIZENBERG, I.; AIZENBERG, N.; VANDEWALLE, J. *Multi-valued and universal binary neurons*: theory, learning and applications. Dordrecht: Springer, 2000.

BIRAN, O.; COTTON, C. Explanation and justification in machine learning: a survey. *Proceedings of IJCAI-17*, Workshop on explainable AI (XAI), 2017.

CADWALLADR, C. Google, democracy and the truth about internet search. *The Guardian*, 4 dec 2016. Disponível em: <http://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook>. Acesso em: 26 maio, 2018.

CLANCEY W. J.; SHORTLIFFE, E. (Orgs.). *Readings in medical artificial intelligence*: the first decade. Reading, MA: Addison Wesley, 1984.

DE SOUZA, C. S. *The semiotic engineering of human-computer interaction*. Cambridge, MA: MIT Press, 2005.

DECHTER, R. Learning while searching in constraint-satisfaction problems. *Proceedings of the 5th National Conference on Artificial Intelligence*. Philadelphia, PA, August 11-15. Vol. 1, p. 178-183, 1986.

DENG, Jia; DONG, Wei; SOCHER, Richard; LI, Li-Jia; LI, Kai; FEI-FEI, Li. Imagenet: a large-scale hierarchical image database. In: *Proceedings of IEEE*, Conference on Computer Vision and Pattern Recognition. p. 248-255, 2009.

DHURANDHAR, Amit; CHEN, Pin-Yu; LUSS, Ronny; TU, Chun-Chen; TING, Paishun; SHANMUGAM, Karthikeyan; DAA, Payel. Explanations based on the missing: towards contrastive explanations with pertinent negatives, 2018. Disponível em: <http://arxiv.org/abs/1802.07623>. Acesso em: 26 mai. 2018.

DORAN, Derek; SCHULZ, Sarah; BESOLD, Tarek. What does explainable ai really mean? a new conceptualization of perspectives. *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*, arXiv:1710.00794, 2017.

ECO, U. On truth: a fiction. In: ECO, U.; SANTAMBROGGIO, M.; VIOLI, P. (Orgs.). *Meaning and mental representations*. Bloomington, IN: Indiana University Press. p. 41-59, 1988.

GOODMAN, B; FLAXMAN, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, vol. 38, no. 3, 2016. Disponível em: <http://doi.org/10.1609/aimag.v38i3.2741>. Acesso em: 26 mai. 2018.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceeding ICCV '15 Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. p. 1026-1034, 2015.

HERLOCKER, J.; KONSTAN, J.; RIEDL, J. Explaining collaborative filtering recommendations. In: *Proceedings of the Third Conference on Computer Supported Cooperative Work* (CSCW), p. 241-250, 2000.

HERN, A. 'Partnership on AI' formed by Google, Facebook, Amazon, IBM and Microsoft". *The Guardian*: International Edition, 29/06/2016. Disponível em: <http://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms>. Acesso em: 4 maio, 2018. Ver também sítio oficial do consórcio: <http://www.partnershiponai.org/>.

LIPTON, Zachary. The mythos of model interpretability. *ICML Workshop on Human Interpretability in Machine Learning*, 2016. Disponível em: <https://arxiv.org/pdf/1606.03490.pdf>. Acesso 17 junho, 2018.

NADIN, M. In folly ripe. In reason rotten: putting machine theology to rest, 2017. Disponível em: <https://arxiv.org/abs/1712.04306v1>. Acesso em: 4 maio, 2018.

PEARL, J. *Probabilistic reasoning in intelligent systems.* San Francisco, CA: Morgan Kaufmann, 1988.

PEIRCE, C. S. *Philosophical writings of Peirce*, Buchler, J. (Org.). New York, NY: Dover Publications, 1955.

SANTAELLA, L. *O método anticartesiano de C. S. Peirce*. São Paulo: Ed. Unesp, 2004.

SEBE, N. Human-centered computing. In: NAKASHIMA, Hideyuki; AGHAJAN, Hamid; AUGUSTO, Juan Carlos (Orgs.). *Handbook of ambient intelligence and smart environments*. Dordrecht: Springer, p. 349-370, 2010.

YOSINSKI, Jason; CLUNE, Jeff; NGUYEN, Anh; FUCHS, Thomas; LIPSON, Hod. Understanding neural networks through deep visualization. In: *Proceedings of the Deep Learning Workshop*, 31st International Conference on Machine Learning, Lille, France. 2015.