A INTELIGÊNCIA ARTIFICIAL COMO METAVOCABULÁRIO PRAGMÁTICO EM ROBERT BRANDOM

ARTIFICIAL INTELLIGENCE AS A PRAGMATIC METAVOCABULARY IN ROBERT BRANDOM

https://doi.org/10.26512/rfmc.v13i2.55078

Ernesto Perini-Santos*
Universidade Federal de Minas Gerais

http://lattes.cnpq.br/2351082986169976
https://orcid.org/0000-0001-5805-5985
eperinisantos@gmail.com

Carlos Barth**
Faculdade Jesuíta de Filosofia e Teologia

http://lattes.cnpq.br/2539678521865135
https://orcid.org/0000-0002-9327-9818
carloshb@protonmail.com

* Professor Titular do Departamento de Filosofia da Universidade Federal de Minas Gerais e Pesquisador 1B do CNPq.
** Pesquisador de pós-doutorado na Faculdade Jesuíta de Filosofia e Teologia (FAJE/MG). Doutor em Filosofia pela Universidade Federal de Minas Gerais.

RESUMO
A Inteligência Artificial clássica tem um lugar fundacional em Brandom: as práticas cujo domínio constitui a posse de um vocabulário são aplicação de uma série de algoritmos. A explicitação desses algoritmos fornece uma explicação para o projeto brandomiano de trazer para o jogo de dar e receber razões os comprometimentos inferenciais implícitos em nossas práticas. Esse projeto falha por uma razão já bem conhecida na IA, o frame problem. Brandom propõe uma solução para o frame problem através de aprendizado por treinamento. A proposta de Brandom se aproxima de redes neurais desenvolvidas através do aprendizado de máquina. Se essa aproximação não permite a manutenção do quadro brandomiano de Between Saying and Doing, o paralelo com Making it Explicit traz uma consequência importante para a compreensão do projeto da Inteligência Artificial Explicável, a saber, tornar explícitos os comprometimentos inferenciais implícitos de processos decisórios que afetam nossa vida em comum.

Palavras-chave: Inteligência artificial. Frame problem. Robert Brandom. Inteligência Artificial Explicável.

Abstract: Classical Artificial Intelligence has a foundational place in Brandom: the practices whose domain constitutes the possession of a vocabulary are the application of a series of algorithms. Making explicit these algorithms provides an explanation for Brandom's project of bringing the inferential commitments implicit in our practices into the game of giving and receiving reasons. This project fails for a reason well known in AI, the frame problem. Brandom proposes a solution to the frame problem through learning by training. Brandom's proposal comes close to neural networks developed through machine learning. While this approach does not allow us to maintain the Brandomian framework of Between Saying and Doing, the parallel with Making it Explicit brings an important parallel with the project of Explainable Artificial Intelligence, namely making explicit the implicit inferential commitments of decision-making processes that affect our common life.

Keywords: Artificial Intelligence. Frame Problem. Robert Brandom. Explainable Artificial Intelligence.

1

> *"So in my terms, classical AI-functionalism claims that computer languages are in principle sufficient pragmatic metavocabularies for some autonomous vocabulary. (Did you see that coming?)"*
> *Robert Brandom, Between Saying and Doing.*

O programa clássico da Inteligência Artificial (IA) tem um papel fundacional no projeto filosófico de Robert Brandom exposto em *Between Saying and Doing* (*BSD*), na confluência de uma virada pragmatista da teoria do significado e da manutenção do programa formalista da tradição analítica. Para compreender como chegamos a essa tese, precisamos dar um passo atrás no desenvolvimento da sua filosofia. Em *Making it explicit* (*MIE*), Brandom apresenta uma teoria pragmatista do significado:

> As expressões passam a significar o que significam ao serem usadas na prática, e os estados e atitudes intencionais têm o conteúdo que têm em virtude do papel que desempenham na economia comportamental daqueles a quem são atribuídos (*MIE*, 134, tradução nossa)[I].

O conjunto de práticas que constituem o significado dos termos, isto é, segundo a semântica inferencialista de Brandom, a rede de inferências nas quais se situam as expressões linguísticas, responde a regras implícitas que são explicitadas no jogo de dar e receber razões. O título do livro designa precisamente a explicitação da estrutura normativa implícita nas nossas práticas. Uma teoria fundacional do significado de expressões linguísticas será uma teoria sobre seu uso, isto é, uma teoria

I Expressions come to mean what they mean by being used as they are in practice, and intentional states and attitudes have the contents they do in virtue of the role they play in the behavioral economy of those to whom they are attributed.

acerca das práticas inferenciais cujo domínio constitui o domínio do significado da expressão ela mesma.

Segundo Brandom, uma teoria pragmatista deve substituir a tese comum na tradição analítica de encontrar na análise filosófica e, em particular, na lógica filosófica, o lugar de tratamento dos problemas filosóficos tradicionais. A versão da teoria pragmatista brandomiana irá permanecer, em certa medida, na continuidade da tradição formalista da filosofia analítica, e é possível que essa seja a razão do papel peculiar que a IA vem a desempenhar no seu projeto teórico: a teoria é constituída por uma apresentação formal das práticas nas quais se engajam agentes que usam um determinado vocabulário, dito de outro modo, dos algoritmos que constituem o uso de um determinado vocabulário, o que nos leva precisamente ao papel central da IA na sua filosofia. Mas estamos avançando no nosso argumento.

Between Saying and Doing (*BSD*), publicado em 2008, parece fornecer um elemento suplementar à teoria proposta por *MIE*. De fato, em *BSD*, Brandom oferece uma teoria para a relação entre práticas e vocabulários que explica o que é, para um metavocabulário pragmático, explicitar as regras que guiam um conjunto de práticas. É precisamente a relação entre uma prática e um metavocabulário pragmático que traz um problema importante para o programa proposto em *BSD*. Para entender como se coloca esse problema, vamos apresentar, na seção 2, os projetos de *MIE* e *BSD* de maneira a realçar sua continuidade. Na seção 3, vamos apresentar o programa clássico da IA e um problema crucial que se coloca para esse quadro teórico, o *frame problem*. A seção 4 é dedicada à solução proposta por Brandom ao *frame problem*. A seção 5 examina uma intrigante associação entre a solução proposta por Brandom a esse problema e o programa contemporâneo da IA, que utiliza redes neurais e técnicas de aprendizagem profunda (*deep learning*). A discussão é aprofundada na seção 6, onde abordamos os recentes esforços para ampliar a interpretabilidade e a explicabilidade das redes neurais. As consequências dessa associação, explorando um paralelo entre a IA e uma observação de Susan Hurley sobre a racionalidade não humana, são o tema da seção 7.

2

Em *MIE*, Brandom explica numa chave expressivista a prática de explicitação de normas implícitas: ao explicitar uma regra seguida implicitamente, o agente expressa os compromissos inferenciais assumidos na sua prática. Dentro do jogo de dar e receber razões, essa é uma maneira de trazer nossas práticas para o controle racional, "*in a form in which they can be confronted with objections and alternatives*" (*MIE*, 106). Vamos chamar a prática inicial, na qual as regras permanecem implícitas, **P1**, e **P2** a prática de explicitação das normas implícitas em **P1**. Se a abordagem expressivista *per se* não se compromete com uma descrição dos mecanismos que constituem as práticas, o desenvolvimento deste projeto em *BSD* não é tão neutro em relação ao modo de funcionamento de **P1** e **P2**.

A explicação proposta em *BSD* para as relações entre as práticas **P1** e **P2** e dessas práticas com os vocabulários para os quais elas são o fundamento resulta numa posição teoricamente mais carregada. A relação de fundamentação entre um vocabulário e uma prática é dita ser **PV-suficiente** para o uso (*deployment*): o agente que se engaja numa prática **P1** sabe o suficiente para dominar um vocabulário **V1**. Há uma demanda suplementar de caracterização de **P1**: **P1** só é teoricamente útil se puder ser especificada num metavocabulário **V2** (*BSD*, 9). A relação entre **V2** e **P1** relação será dita **VP-suficiente** para caracterização. A relação entre **V1** e **V2** é derivada das relações **PV-suficiente** e **VP-suficiente**. O primeiro esquema é o seguinte:

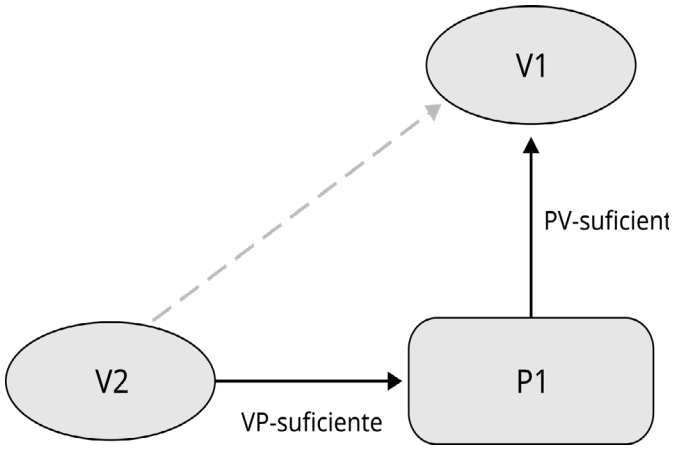

Figura 1: *Relações entre V*1, *P*1 *e V*2

Falta um componente teórico: se todo vocabulário deve ser baseado numa prática, devemos perguntar a que prática corresponde **V2**. A ameaça de simplesmente postular uma relação análoga a **P1-V1** não é apenas de nos engajarmos num regresso sem um termo, mas também de não explicar a relação **VP-suficiente** entre **V2** e **P1**. A solução de Brandom consiste em postular uma prática **P2** que está na relação **PV--suficiente** com **V2**, mas que será também uma reelaboração algorítmica de **P1**. Um esquema das relações entre práticas e vocabulários deve então ser desenvolvido do seguinte modo:

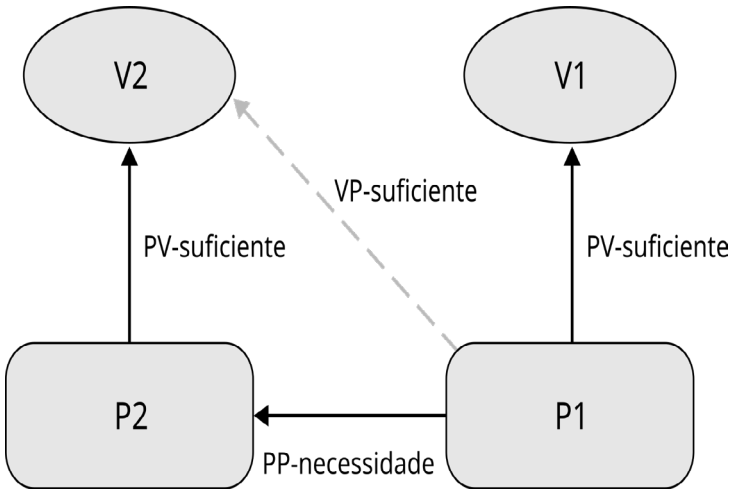

Figura 2: *Relações entre práticas e vocabulários*

P2 envolve os mesmos algoritmos que **P1**, mas reorganizados de uma maneira que eles não são mais capazes de fazer o que o agente faz ao se engajar em **P1**. **P2** é a fundamentação do metavocabulário **V2** que especifica **P1**, mas não fornece uma substituição de **V1**: **V2** é expressivamente mais fraco do que **V1**. A especificação não é outra coisa senão a explicitação dos algoritmos que constituem **P1**.

Recapitulando o que vimos até aqui, as relações entre práticas e vocabulários são as seguintes:

> a) Uma prática **P1** é suficiente para fundamentar o uso de um vocabulário **V1** – trata-se de uma relação de **PV-suficiência** de uso.
> b) A prática **P1** deve ser descrita por um vocabulário metapragmático **V2**, que irá especificar as práticas-e-habilidades que constituem **P1** – uma relação de **VP-suficiência** de especificação;
> c) O vocabulário **V2** deve corresponder também a uma prática **P2**, numa relação de **PV-suficiência**. **P2** é a reelaboração algorítmica das práticas-e-habilidades que constituem **P1** – uma relação **PP-necessária** de dependência entre dois conjuntos de práticas (*BSD*, 13).

Utilizando os termos de *MIE*, **V2** torna explícito o padrão normativo que está implícito em **P1**. Aqui, nós encontramos um primeiro problema, ou pelo menos uma primeira bifurcação: a descrição das relações entre práticas e vocabulários em *BSD* sugere que **V2** representa o ponto de visa teórico, que especifica os comprometimentos que constituem o uso de **V1** ao tornar explícitos os algoritmos que constituem **P1**. De fato, apenas do ponto de vista do teórico faz sentido esperar a especificação de algoritmos que correspondem a uma prática determinada. Ora, *MIE* parece exigir um engajamento dos agentes eles mesmos, no ajuste mútuo resultante de trocas conversacionais, de maneira que a explicitação que aparece no jogo de dar e receber razões representa o ponto de vista do agente. De resto, o ponto de vista do agente parece mais compatível com a leitura expressivista da lógica.

Se, para relações inferenciais mais simples, estes dois pontos de vistas parecem convergir, mais ou menos na linha de um célebre artigo de Gilbert Ryle (1950), a capacidade de representação, pelo agente, do que constitui sua prática é menos evidente. Essa tensão foi aponta por Kevin Scharp, já em *MIE*: a lista de comprometimentos atribuídos a um sujeito, isto é, a explicitação da estrutura normativa implícita de suas práticas, é feita pelo teórico, não pelo agente ele mesmo (Scharp, 2005, p. 211–212). Ora, o jogo de dar e receber razões no qual se engajam os agentes deve recorrer ao que é acessível aos agentes eles mesmos, e não a um vocabulário disponível para um teórico. Esse conflito mostra a atenção que se deve ter ao ponto de vista a partir do qual um dado vocabulário é disponível – se é que há um vocabulário disponível – para a especificação de uma prática. É bem possível, com efeito, que a especificação dos compromissos inferenciais explicitados no jogo de dar e receber razões não recorra ao mesmo vocabulário, nem seja a mesma prática, da recombinação de algoritmos. Interessa-nos aqui um outro problema que diz respeito à existência mesma de um vocabulário **V2**, ou pelo menos de um vocabulário **V2** com a profundidade necessária para desempenhar o papel que ele deve ter no esquema brandomiano de *BSD*.

De fato, o que há de mais revelador aqui vem da relação entre práticas e vocabulários como uma relação de reelaboração algorítmica. A tese se desdobra em duas:

- uma prática **P1** suficiente para o domínio de um vocabulário **V1** autônomo pode ser decomposta num conjunto de algoritmos mais básicos;
- esses algoritmos podem, por sua vez, ser recompostos em uma outra prática **P2**, que levará a um outro vocabulário **V2** que é especificação dos algoritmos constitutivos a **P1**. [II]

II *"What I will call the "algorithmic pragmatic elaboration" version of AI-functionalism—or just "pragmatic AI"—is the claim that there is a set of practices-or-abilities meeting two conditions*: *It can be algorithmically elaborated into* (*the ability to engage in*) *an autonomous discursive practice* (*ADP*). *Every element in that set of primitive practices-or-abilities can intelligibly be understood to be engaged in, possessed, exercised, or exhibited by something that does not engage in any ADP*" (*BSD*, 75).

Com essas duas teses em vista, compreendemos o metavocabulário pragmático **V2** como uma descrição algorítmica de **P1**. Ora, propor um vocabulário teórico que especifica os algoritmos que constituem uma determinada prática é o que faz a inteligência artificial. A IA clássica passa assim a ter um papel central no projeto de Brandom:

> E isso quer dizer que uma linguagem de computador, na qual qualquer algoritmo desse tipo pode ser expresso, é, em princípio, VP-suficiente para especificar habilidades que são VP-suficientes para usar um vocabulário autônomo. **Então, em meus termos, o funcionalismo clássico da IA afirma que as linguagens de computador são, em princípio, metavocabulários pragmáticos suficientes para algum vocabulário autônomo.** (Você esperava por essa?) (*BSD*, 70, tradução nossa)[III].

É por esta razão que, de maneira talvez surpreendente (ou pelo menos que Brandom acha surpreendente), a IA toma o papel da lógica na fundamentação da filosofia da linguagem. De fato, a tese de Brandom supõe que práticas que sustentam inferências usuais são algoritmos que podem ser descritos num metavocabulário. Não é certo, contudo, que seu projeto possa ser levado a um bom termo.

3

Brandom busca na IA clássica um modo de especificar metavocabulários pragmáticos para práticas usuais. Para avaliar a viabilidade dessa proposta, precisamos antes resgatar alguns elementos centrais da IA.

III *And that is to say that a computer language, in which any such algorithm can be expressed, is in principle VP-sufficient to specify abilities that are PV-sufficient to deploy an autonomous vocabulary. So in my terms, classical AI-functionalism claims that computer languages are in principle sufficient pragmatic metavocabularies for some autonomous vocabulary*. (*Did you see that coming?*)

Inicialmente, devemos nos lembrar que a IA envolve empreitadas com concepções de inteligência distintas. Alguns pesquisadores se interessam pela inteligência enquanto um fenômeno humano. Não basta simular comportamentos inteligentes, é preciso que esses comportamentos se originem em processos e mecanismos estruturalmente semelhantes aos encontrados na cognição humana. Contudo, há linhas de pesquisa que tomam a inteligência como um fenômeno mais amplo, e o modo como ela se manifesta em seres humanos é apenas um dentre outros possíveis. Para estes pesquisadores, o que interessa é o comportamento resultante do sistema. Não importa, por exemplo, se os algoritmos usados forem biologicamente implausíveis ou se as demandas de poder computacional e memória excederem as nossas, se eles levarem a resultados que, para seres humanos, manifestam nossa capacidade de pensar. Em outras palavras, a IA acomoda a possibilidade de agentes inteligentes inumanos.

Em razão desta duplicidade de projetos, a palavra "agente" tem na IA um sentido distinto daquele tipicamente associado ao termo, particularmente em filosofia. Na perspectiva da IA, um agente é apenas um sistema que recebe informações do seu ambiente (que pode ser virtual) e reage de forma adaptativa. No entanto, para a extensão dessa distinção ao projeto de Brandom, devemos aqui considerar somente agentes humanos. Afinal de contas, para que a IA funcione como um vocabulário metapragmático para práticas humanas, somente modelos computacionais que se preocupam com os constrangimentos de ordem biológica e psicológica têm alguma serventia.

Como vimos acima, no projeto de Brandom, ou pelo menos na convergência dos projetos de *MIE* e *BSD*, há uma tensão acerca do ponto de vista que corresponde ao uso de um vocabulário metapragmático. Aqui, vamos considerar o ponto de vista do teórico, isto é, nos termos da IA, do desenvolvedor. Essa escolha parece, de fato, ser a única compatível com o funcionamento de **V2** como uma especificação dos algoritmos que constituem **P1**, que é precisamente o que explica papel reservado à IA em *BSD*.

O programa clássico da IA tem como principal característica a modelagem de capacidades cognitivas por meio de modelos computacionais clássicos (Samuels, 2018). Esse tipo de modelo tem propriedades metodológicas relevantes para o projeto de Brandom, e isso pode nos ajudar a entender sua opção em *BSD*. Eles envolvem grande número de inferências em série, que se realizam sobre uma quantidade relativamente pequena de variáveis fortemente correlacionadas (causalidade, por exemplo) (Smith, 2019). Isso faz com que sejam interpretáveis, isto é, podemos dar sentido ao algoritmo e descrevê-lo como uma sucessão de estados dotados de significado.[IV] Além disso, a modelagem clássica requer que o desenvolvedor seja explícito acerca de todas as suposições e todas as etapas envolvidas, o que evita modelos que pressupõem subrepticiamente as capacidades que almejam explicar.

Esse é o tipo de característica que fez emergir o agora famoso *frame problem* (McCarthy; Hayes, 1969). Trata-se do problema de como recortar apenas o que é contextualmente relevante para a realização de uma tarefa e ignorar o restante. Fazer esse recorte é enquadrar a situação em um dado *frame*. Isso está por trás da nossa capacidade de pensar e agir de um modo adaptativo e sensível às circunstâncias. *Frames* são necessários porque nossas crenças, desejos e expectativas podem ser integradas e articuladas de múltiplas formas, em um número indefinidamente multiplicável de contextos. O conjunto de articulações possíveis não é fruto - apenas - dos compromissos inferenciais associados a cada crença, mas também das possíveis interdependências entre elas. Isso leva a uma explosão combinatorial que inviabiliza a consideração exaustiva de todas as possíveis articulações. O que evita a explosão combinatorial é a escolha de um *frame*, que seleciona apenas as inferências que são relevantes num dado contexto. No entanto, ao mesmo tempo em

IV Embora a semântica desses modelos costume fazer uso da noção de símbolos, Brandom nota que esta não é a única possibilidade: "*AI functionalism traditionally held itself hostage to a commitment to the purely symbolic character of intelligence in the sense of sapience. But broadening our concern from automata as purely syntactic engines to the realm of transducing automata puts us in a position to see AI functionalism as properly concerned with the algorithmic decomposability of discursive* (*that is, vocabulary-deploying*) *practices--and-abilities*" (*BSD*, 27). Para uma discussão sobre a relação entre conteúdo simbólico e mecanismos computacionais, ver também Piccinini (2015).

que agentes humanos evitam a explosão combinatorial ao restringir as informações relevantes num dado curso de ações, eles também são capazes de reconhecer contextos em que é necessário considerar outras informações, isto é, são capazes de sair das restrições impostas por um dado *frame*, por exemplo, ao continuar sua ação diante de informações novas não previstas inicialmente. O *frame problem* reside na dificuldade de se explicar essas duas habilidades correlatas.[V]

Nessa descrição, o termo "frame" designa um *explanandum*: a capacidade humana de enquadrar o contexto atual de modo a tornar saliente o que é circunstancialmente relevante. Contudo, em alguns casos o termo foi utilizado para referenciar um *explanans*, i.e. uma estratégia usada para explicar essa capacidade. Minsky (1997), por exemplo, utilizou o termo "frame" para designar estruturas de dados que modelam de objetos a situações e perspectivas. O mesmo acontece no trabalho de McCarthy; Hayes (1969), onde o termo caracteriza um tipo de axioma lógico que delimitaria os efeitos de uma ação ou evento.[VI] Estas são diferentes formas de modelar os tipos de situação em que o agente pode se encontrar. A cada tipo de situação corresponde uma estrutura de dados que explicita as regras das práticas usuais e os fatores relevantes para um raciocínio adequado. Os tipos delimitam as inferências disponíveis, funcionando como roteiros ou *scripts*.[VII] Um sistema que selecione o *script* adequado à sua situação corrente é capaz de evitar a explosão combinatorial, ao ignorar todas as regras e fatores não explicitados no *script*. De outro modo: *scripts* são uma forma de modelar *frames*. A capacidade humana de enquadrar a situação corrente num *frame* é explicada pela capacidade que seus mecanismos cognitivos têm de selecionar e seguir um *script*.

V Para uma discussão mais aprofundada, ver Barth (2019, 2024).

VI Trata-se de um formalismo lógico chamado cálculo de situação (*situation calculus*). Uma descrição dessa abordagem pode ser encontrada em Shanahan (1997).

VII O termo "script" foi usado por Schank; Abelson (1977) para designar um tipo específico de estrutura de dados. Contudo, aqui o termo é usado num sentido mais amplo: ele abarca qualquer estratégia usada para delimitar o conjunto de práticas e fatores numa dada situação.

Infelizmente, esse tipo de abordagem é insuficiente para explicar a capacidade humana de agir em situações cambiantes e imprevisíveis, e por isso ela deixa o *frame problem* ileso. O que fazemos no contexto de uma situação de tipo t não depende apenas daquilo que sabemos sobre situações de tipo t, mas também de nossa capacidade de responder a informações que não estão previstas no *script* de t. Além disso, *scripts* não iluminam a capacidade de determinar o tipo de situação em que nos encontramos, mas antes pressupõem que isso já foi feito em etapa anterior.

Como exemplo, suponha-se que queiramos desenvolver um *script* para guiar o comportamento de um agente que se encontra na situação DOIS AMIGOS NUM BAR. Para isso, observaremos cenários possíveis dentro dessa situação e formularemos regras que permitirão determinar o comportamento adequado em cada caso. Comecemos com três cenários simples:

1. Alan e Beto são colegas e estão num bar num fim de tarde. Alan diz a Beto: vamos beber uma cerveja?
2. Alan e Beto são colegas e estão num bar num fim da tarde. Eles vão voltar ao trabalho depois do *happy hour*. Apesar de querer beber uma cerveja, Alan decide não sugerir isto ao amigo.
3. Alan e Beto são colegas e estão num bar num fim de tarde. Ambos vão para casa descansar depois do bar. Beto tem um antigo problema de alcoolismo e diz a Alan que está há dois meses sem beber. Apesar de querer beber uma cerveja, Alan decide não sugerir isto ao amigo.

Um *script* que poderia dar conta dos cenários (1)-(3) demandaria, no mínimo, o seguinte conjunto de regras, articuladas na seguinte ordem: [VIII]

a) SE (A e B são colegas) → (A pode convidar B a beber uma cerveja)
b) SE (A vai trabalhar) → (A não deve beber uma cerveja)

VIII Essa forma de apresentação de *scripts* é vagamente inspirada na arquitetura de sistemas de produção, fortemente associadas ao trabalho de A. Newell. Foi necessário sacrificar acurácia e rigor técnico para tornar as ideias centrais mais acessíveis.

c) SE (A vai descansar) → (A pode beber uma cerveja)
d) SE (B foi alcoólatra) → (A não deve beber, nem convidar B a beber uma cerveja)

Cada regra especifica uma condição de satisfação e uma ação resultante. Elas são analisadas na ordem de apresentação. Assim, quando duas regras – cujas condições tenham sido satisfeitas - se contradizem, a ação resultante será determinada pela sequência de análise.

Considere por exemplo, o resultado da aplicação deste *script* ao cenário (1). A regra (a) é satisfeita: Alan e Beto são colegas, e portanto Alan pode convidar Beto a beber uma cerveja. As regras (b), (c) e (d) são ignoradas porque não são satisfeitas, uma vez que as informações com as quais elas trabalham estão ausentes em (1). No cenário (2), tanto a regra (a) quanto a regra (b) são satisfeitas, fazendo com que Alan desista de beber uma cerveja e, consequentemente, de convidar Beto. Por fim, no cenário (3) a única regra não satisfeita é (b), fazendo com que, ao final da aplicação do *script*, Alan decida não convidar Beto a beber uma cerveja. Como se vê, as regras que efetivamente guiam o comportamento do agente variam em função de como a situação é caracterizada. Como Alan pode decidir se sugerir a Beto beber uma cerveja é adequado ou não? Enquanto o cenário (1) parece apresentar uma variação típica, na medida em que incluímos mais informações, a mera adequação da pergunta parece depender de outros fatores que devem ser acrescentados ao *script* da situação DOIS AMIGOS NUM BAR. Tudo se complica pelo fato de sempre podermos acrescentar ainda mais informações:

4. Alan e Beto são colegas e estão num bar num fim da tarde. Eles vão voltar ao trabalho depois do *happy hour*. Apesar de querer beber uma cerveja, Alan decide não sugerir isto ao amigo. Em seguida, Alan pensa que o trabalho não é tão complicado e que uma cervejinha não fará mal. Alan diz a Beto: você quer beber uma cerveja?

5. Alan e Beto são colegas e estão num bar num fim de tarde. Ambos vão para casa descansar depois do bar. Beto tem um antigo problema de alcoolismo e diz a Alan que está há dois

> meses sem beber. Apesar de querer beber uma cerveja, Alan decide não sugerir isto ao amigo. Beto sabe que Alan quer beber uma cerveja. Ele diz que pode perfeitamente pedir uma água enquanto seu amigo satisfaz sua vontade.

O *script* apresentado acima seria incapaz de explicar o comportamento de Alan nos cenários (4) e (5). Ele não possui nenhuma regra especificando o que é adequado fazer quando o trabalho a ser feito depois de beber uma cerveja não é complicado. O mesmo acontece no cenário (5): o *script* não especifica o que deve ser feito com a informação de que Beto não vê problema em Alan beber uma cerveja sozinho. É preciso acrescentar regras adicionais, (e) e (f), para que essas variações sejam acomodadas:

> a) SE (A e B são colegas) (A pode convidar B a beber uma cerveja)
> b) SE (A vai trabalhar) (A não deve beber uma cerveja)
> c) SE (A vai descansar) (A pode beber uma cerveja)
> d) SE (B foi alcoólatra) (A não deve beber, nem convidar B a beber uma cerveja)
> e) SE (A vai trabalhar com algo não complicado) (A pode beber uma cerveja e convidar B a beber uma cerveja)
> f) SE (B não se importar) (A pode beber uma cerveja)

O problema que se mostra é que não parece haver um limite para o número de fatores e de regras que regem o comportamento adequado mesmo em situações simples como a de DOIS AMIGOS NUM BAR. Alan e Beto são colegas ou também amigos? Há chance de algum outro colega de trabalho aparecer? É uma data especial? Algum deles está na expectativa de uma notícia importante? Inúmeros fatores como estes podem influenciar a decisão de Alan. Mesmo em cenários artificialmente simplificados como os descritos acima, a complexidade do *script* pode se tornar rapidamente proibitiva. Conforme nos aproximamos da riqueza informacional característica das atividades humanas, o resultado é uma explosão combinatorial que é, ao mesmo tempo, inescapável e intratável.

Esse crescimento rápido e desenfreado da complexidade não se deve apenas à adição de um número ilimitado de novos fatores. As diferen-

tes maneiras de se pensar numa situação (do ponto de vista do agente), ou as diferentes maneiras de se caracterizar uma situação (do ponto de vista do teórico), já refletem escolhas e interesses anteriores à escolha mesma do *script*. Isso significa que há um número indefinidamente multiplicável de articulações dos fatores já dados numa determinada caracterização. Como exemplo, considere a seguinte situação:

6. Alguns amigos vão a um restaurante para comer uma moqueca de camarão, que, no entanto, está em falta. Eles irão comer um outro prato, ou mudam de restaurante?

Esta situação pode ser descrita de dois modos, ou cair sob dois tipos diferentes:

7. Alguns amigos vão jantar fora para comemorar o aniversário de um deles. Eles pretendem presenteá-lo com seu prato favorito, moqueca de camarão.
8. Alguns amigos vão jantar fora para comemorar o aniversário de um deles. Eles pretendem presenteá-lo pagando a conta.

Na descrição (7), parece mais adequado mudar de restaurante, mas não na situação (8). Qual o tipo mais adequado para caracterizar a situação? Deve o *script* que seleciona as regras a guiar o comportamento dos envolvidos levar em conta a natureza do presente? Se não, precisaremos de um *script* para uma situação do tipo "*comemorar um aniversário com amigos num restaurante*". Se sim, precisaremos de um *script* para uma situação do tipo "*comemorar um aniversário com amigos num restaurante em que o aniversariante será presenteado com seu prato favorito*". A princípio, isso pode parecer excessivamente específico para tipificar uma situação. No entanto, a decisão sobre qual é o comportamento adequado na situação (6) depende exatamente desta escolha.

Isso mostra que um *script* que não articule todos os elementos necessários resultará numa subdeterminação do comportamento adequado. Porém, não há como prever o tipo de elemento que pode afetar a decisão sobre o enquadramento adequado de uma situação (i.e. qual *frame*

utilizar). Esse problema parece inescapável, pois o conjunto de possíveis formas de categorizar uma situação é indefinidamente multiplicável.

Em síntese, a seleção de regras depende da seleção do *script*. Mas como selecionar o *script* adequado? Um algoritmo de seleção de *scripts* que se pretenda geral demandaria a consideração exaustiva de todas as possíveis situações - justamente o que se pretendia evitar. Qualquer algoritmo que trabalhe com suposições contextuais, isto é, que evite a explosão combinatorial, só será funcional no interior do seu contexto alvo, ou seja, ele irá pressupor que está sendo aplicado numa situação em que as suposições que o caracterizam coincidem com as que são adequadas à situação corrente. Isso supõe - em vez de fornecer - uma solução ao *frame problem*, pois é preciso explicar como se dá a formulação e a seleção desse *frame* contextualmente adequado. O que o *frame problem* nos mostra, portanto, é que não há algoritmo para a seleção de *scripts*, ou seja, não há algoritmo para a seleção de *frames*.

4

Um teórico que buscasse caracterizar o comportamento de Alan e Beto no cenário (1) deveria propor algoritmos nos quais as práticas de sair do trabalho e ir a um bar beber uma cerveja seriam decompostas em algoritmos mais básicos. Seja assim **P1** a prática de ir um bar com um amigo e **V1**, o vocabulário usado nesta situação. É preciso haver um **V2** que caracterize a prática **P1**, explicando a decisão de Alan de propor a Beto beber uma cerveja. No entanto, **V2** deveria ser capaz de cobrir não apenas o cenário (1), mas também as variações (2)-(5) e todas as demais possíveis. Ou seja, **V2** precisaria ser capaz de acomodar não apenas descrições parciais como as aqui fornecidas, mas um número indefinidamente multiplicável de fatores e um número igualmente aberto de modos pelos quais eles poderiam participar da situação. Dito de outro modo, **P1** deve comportar algoritmos que cubram os cenários (2)-(5), assim como todas os possíveis acréscimos informacionais que poderiam afetar o comportamento dos agentes, e esses algoritmos devem ser es-

pecificados por **V2**. Contudo, o *frame* problem implica que não há um tal vocabulário. Ele nos mostra que não há algoritmos que constituem uma prática **P1** para a transição entre *frames* como os descritos acima. Não havendo algoritmos que constituem a extensão indefinida de **P1**, também não há **V2**. Segue-se que a análise de Brandom não se aplica a situações que podem exigir dos agentes a adequação a informações não previstas, isto é, toda ação humana.

Brandom parece aceitar que o *frame problem* impõe um limite à sua abordagem. Ele busca mitigar seus efeitos caracterizando uma relação de **PP-suficiência** alternativa: a elaboração prática por meio de treinamento.[IX] Práticas que se mostrem resistentes à análise algorítmica seriam melhor abarcadas por essa relação. Em vez de acomodar a decomposição em algoritmos mais simples, nela temos regimes de treinamento decomponíveis em (grosso modo) práticas pedagógicas. Aqui, uma precisão terminológica parece necessária. Há dois tipos de relação entre práticas que são amalgamadas, a nosso ver erroneamente, por Brandom. A primeira relação de **PP-suficiência** é uma relação de reelaboração algorítmica, que leva a uma prática **P2** e a um vocabulário **V2** metapragmático, que é expressivamente mais pobre do que vocabulário alvo **V1**. Ora, no tratamento do *frame problem*, Brandom postula uma relação entre práticas do mesmo nível – nos exemplos de Brandom, as práticas de desenhar e de realizar outras tarefas que exigem o controle motor fino e de somar e multiplicar (*BSD*, 84-88). Nenhuma destas práticas está num nível metadiscursivo em relação à outra. Como esta distinção será importante para o desenrolar do nosso argumento, vamos distinguir **PP-suficiência**, a relação metapragmática entre **P1** e **P2** e, de maneira derivada, entre **V1** e **V2**, e a relação **PP'-extensão**, entre práticas **P1** e **P1'** que se encontram no mesmo nível. Os vocabulários **V** e **V'** não estão numa relação de especificação, mesmo se **P1'** pode ser

IX A estratégia pode ser associada, ainda que modo vago, às críticas de Dreyfus à IA clássica (1992). Embora a abordagem de Dreyfus envolvesse preocupações com a linguagem, ela se desdobrava no que Cupani (2011) denomina "filosofia fenomenológica da tecnologia". Ainda que por outros caminhos, Dreyfus também percebeu a enorme dificuldade em fornecer descrições algorítmicas dos contexto de atividade e da sensibilidade à relevância.

vista como a reelaboração de **P1** (essa pode ser, de resto, a origem da confusão de Brandom).

A natureza e o alcance dessas práticas de extensão, contudo, permanecem vagos e potencialmente opacos. Brandom reconhece que elas podem se sujeitar a fatores empíricos de ordem biológica, psicológica e mesmo traços individuais do educador e do educando, mas sua caracterização não vai muito além de tomá-las como não algorítmicas. Talvez por isso Brandom não tenha a pretensão de especificar o que exatamente nesses regimes de treinamento permite a superação do *frame problem*. Ele antes supõe que algo de caráter não algorítmico seja necessário, e que isso deve estar presente nesses regimes de treinamento.[X] Sobre a importância da elaboração prática, ele nos diz:

> Penso que a apreciação da centralidade desse tipo de relação de PP-suficiência - que ocorre quando, de fato, criaturas de um certo tipo que podem se engajar em uma prática (exibir uma habilidade) podem ser levadas ou podem aprender a se engajar em (ou exibir) outra - é uma das ideias mestras que animam o pensamento do Wittgenstein tardio. Repetidamente, ele enfatiza o quanto nossas práticas discursivas são possibilitadas pelo fato de que, por conta de um fato contingente, aqueles que têm um conjunto de habilidades ou podem se engajar em um conjunto de práticas podem ser leva-

X Note que esta suposição é muito mais forte do que pode parecer a princípio. A natureza do *frame problem* é objeto de disputa desde seu surgimento. Contemporaneamente, poucos pesquisadores dedicados ao tema acreditam que ele emerja apenas em modelos computacionais como os utilizados na IA clássica. Conforme Dennett: *It apparently arises from some very widely held and innocuous-seeming assumptions about the nature of intelligence, the truth of the most undoctrinaire brand of physicalism, and the conviction that it must be possible to explain how we think* (Dennett, 1987, p. 44).

> dos pelo treinamento a exibir ou se engajar em outro (*BSD*, p. 85, tradução nossa)[XI].

Do nosso ponto de vista, a preocupação mais aguda para Brandom não é especificar se e como uma relação de **PP-extensão** ancorada no treinamento consegue evitar o *frame problem*. Para os propósitos de nossa argumentação, podemos inclusive assumir que esse é o caso. O problema maior é que são precisamente contextos como os exemplificados nas situações (1)-(5) que parecem demandar a mediação de um metavocabulário **V2** que que funcione como um modulador entre uma situação **P1** e uma situação **P1'** que estejam numa relação de transição entre *frames*. Dito de outro modo, é neste tipo de contexto que a prática da explicitação parece fazer sentido. Mas se a prática que explica a alternância entre **P1** e **P1'** for ancorada em treinamento, ela não é coberta pelo que Brandom diz ser a explicação via reelaboração algorítmica. O recurso a um metavocabulário que explicite as normas implícitas parece ser especialmente útil na transição entre *frames*, precisamente quando não há reelaboração algorítmica possível. Se recorremos à relação entre *MIE* e *BSD*, parece que precisamente nos casos em que o jogo de dar receber e receber razões faz mais sentido, o esquema delineado em *BSD* é inaplicável.

Se esse resultado pode parecer com um interesse apenas interno à filosofia de Brandom – talvez fosse suficiente desvincular os projetos de *MIE* e *BSD* –, a falha na transição algorítmica pode revelar algo mais profundo. Podemos, de fato, imaginar que, dentro de cada *frame*, o esquema de Brandom funcione e talvez mesmo explique ajuste mais locais, mas que a transição entre *frames* se dê uma maneira não tratável pela via da reelaboração algorítmica. Como se houvesse ilhas de transparência interpretativa num oceano de práticas que navegamos de ma-

XI *I think an appreciation of the centrality of this sort of PP-sufficiency relation - which obtains when, as a matter of fact, creatures of a certain sort who can engage in a practice (exhibit an ability) can be brought or can learn to engage in (or exhibit) another - is one of the master ideas animating the thought of the later Wittgenstein. Again and again he emphasizes the extent to which our discursive practices are made possible by the fact that, as a matter of contingent fact, those who have one set of abilities or can engage in one set of practices can be brought by training to exhibit or engage in another.*

neiras que nos restam opacas, isto é, práticas para as quais somos treinados. Esta tese não quer dizer que a transição entre *frames* não seja guiada por algoritmos, mas que eles não são completamente especificáveis – é bem possível que esta cartografia represente um mapeamento do que podemos conhecer. Mas mesmo lido numa chave epistêmica, esta tese tem consequências importantes para o projeto de Brandom, afinal de contas, **V2** não é outra coisa senão uma especificação dos algoritmos que constituem **P1**.

A ideia de ilhas algorítmicas deve nos remeter aqui à tese de Susan Hurley de que as capacidades de raciocínio de animais não humanos constituem "ilhas de racionalidade, mais do que um espaço contínuo de razões" (Hurley, 2006, p. 139). Esta tese poderia ser expandida de dois modos, quer postulando uma racionalidade mais local para humanos (talvez nós apenas habitemos ilhas um pouco maiores), quer explicando o espaço contínuo de razões através de mecanismos heterogêneos. Voltaremos a esse ponto.

5

A ampla adoção de modelos neurais é uma das marcas da IA contemporânea. Como se sabe, esse tipo de modelo não é construído manualmente, mas antes gerado pela aplicação de algoritmos de aprendizagem de máquina (*machine learning* ou *ML*) a um dado conjunto de informações, que constitui seu *corpus* de treinamento. Embora tenha se tornado comum utilizar a expressão ML para se referir a algoritmos que geram modelos neurais, em um sentido mais rigoroso, o ML também acomoda técnicas de aprendizagem que geram outros tipos de modelo. Alguns deles têm estrutura compatível com o que se esperaria encontrar na IA clássica, tais como árvores de decisão. A hipótese aqui investigada, contudo, é a de que há um paralelo entre a transição não plenamente especificável entre diferentes práticas de base **P1/P1'** e redes neurais. Essa delimitação se justifica não apenas pela prioridade que as pesquisas contemporâneas dão a esse tipo de modelo, mas também a proprieda-

des como multidimensionalidade (representação em múltiplas dimensões interligadas) e superposição (combinação simultânea de múltiplos elementos). Embora não sejam exclusivas a esse tipo de modelo, redes neurais permitem que essas características sejam exploradas de maneira muito mais escalável, flexível e eficiente. Isso faz com que o ML focado em redes neurais se aproxime mais da ideia de Brandom para a transição entre *frames* por meio do treinamento, pois como veremos, a forma pela qual o aprendizado emerge em modelos neurais é radicalmente distinta da preconizada pela reelaboração algorítmica. Com efeito, no que se segue utilizaremos a expressão "ML" para nos referirmos ao treinamento de modelos neurais.

O objetivo do algoritmo de treinamento é identificar padrões presentes no *corpus*, na esperança de que o modelo resultante consiga reconhecer esses padrões mesmo ao lidar com variações originalmente indisponíveis nos dados de treinamento. Ou seja, é esperado que o modelo adquira alguma capacidade de generalização, e que possa fazer uso disso ao decidir como agir em situações semelhantes àquelas a partir das quais seu treinamento se deu. Modelos neurais são capazes disso porque tendem a capturar padrões sutis e de difícil detecção em um número massivo de variáveis. Além disso, na maioria dos casos a covariação mapeada entre essas variáveis é significativamente fraca (e.g. a presença de uma propriedade *a* pode aumentar em 1.3% a chance de a propriedade *b* também estar presente). Isso significa que, diante de um *corpus* de treinamento que apresente um grande número de variações da situação DOIS AMIGOS NUM BAR, o modelo resultante pode ser capaz de selecionar entre práticas **P1** e **P1'** por caminhos de difícil compreensão. Nesse sentido, é possível dizer que modelos neurais bem treinados tendem a capturar um maior número de permutações de uma mesma situação do que um desenvolvedor seria capaz de especificar por conta própria. Esse número de casos tende a aumentar rapidamente, conforme o volume de informações disponibilizadas no *corpus* de treinamento se amplia. Com efeito, modelos neurais podem entregar uma considerável expansão no número de casos cobertos.

Não poderia a adoção de modelos neurais constituir uma solução para o *frame problem*? Isso já foi amplamente investigado, e a conclusão sem-

pre foi negativa.[XII] Em modelos neurais, o *frame problem* toma a forma de um problema de roteamento: como definir quando, e em que circunstâncias, certas subpopulações neurais terão acesso à informação sendo processada? De modo mais sintético: como modular o fluxo informacional? O modo como a informação flui ao longo da rede neural leva o modelo a salientar certos padrões (*features*) e mitigar outros. A decisão sobre como rotear a informação é, portanto, análoga à decisão sobre como uma determinada situação deve ser enquadrada, i.e. sob qual *frame* ela deve ser considerada. Na IA contemporânea, esse problema se mostra como o desafio de extrapolar para além da distribuição probabilística dos padrões presentes no *corpus* de treinamento (*out-of-distribution extrapolation*) (Arjovsky, 2021; Liu et al., 2021). Por exemplo, alguns modelos "aprendem" que a textura da pele de elefantes é mais importante do que a forma ao identificar elefantes (Geirhos et al., 2019). Isso pode resultar na capacidade de reconhecer elefantes em certos contextos, como quando a textura de sua pele está claramente visível, mas na impossibilidade de detectá-los quando as condições de iluminação deixam apenas sua silhueta visível. O que era para ser um detector de elefantes termina sendo um detector de pele de elefantes. Na ausência de vieses cuidadosamente desenhados pelos desenvolvedor ("nesse caso, forma tem peso maior que textura"), o algoritmo de treinamento não consegue decidir de modo fluido e adaptativo quando uma determinada propriedade é relevante. O problema, claro, é que tais vieses constituem soluções locais *ad hoc* para problemas globais. Não por acaso, os modelos existentes costumam se ater a domínios específicos.[XIII]

Se nossa hipótese estiver correta e a solução do *frame problem* pela aprendizagem puder ser compreendida nos termos do ML, então Bran-

XII Embora Churchland (1989) tenha afirmado que modelos neurais não sofrem do *frame problem*, isso é demonstravelmente falso. Para uma resposta direta a Churchland, ver Clark (2002). Para detalhes adicionais, ver Haselager; Rappard (1998) e Barth (2021).

XIII Isso vale também para grandes modelos linguísticos como os da série GPT produzidos pela OpenAI. Eles mapeiam e exploram a distribuição probabilística de elementos sintáticos, mas ignoram o domínio de conhecimento sobre o qual aqueles elementos versam.

dom não traz uma solução satisfatória para o problema. O lado positivo, contudo, é que por manter o caráter algorítmico da relação de **PP-suficiência**, o projeto tende a se beneficiar de qualquer eventual avanço rumo a uma solução no âmbito da IA. Mesmo mantendo um certo ceticismo em relação à solução do *frame problem* por redes neurais, podemos considerar a expansão de casos tratados como um ganho significativo, o que daria um maior alcance à tese de Brandom acerca da extensão de práticas via treinamento.

Nesse cenário de expansão dos casos cobertos pela IA, a relação entre **P1** e **P1'** e, consequentemente, a relação de **PP-suficiência** entre uma prática de base **P1** e uma prática metapragmática **P2**, passam a ser uma espécie de amálgama entre a elaboração prática e a elaboração algorítmica. Com efeito, a transição entre **P1** e **P1'** partilha do caráter parcialmente opaco da articulação prática. Essa afirmação se justifica por duas razões. Num sentido estrito, um modelo neural treinado é constituído por um conjunto de vetores ou matrizes de "pesos" e "vieses", i.e. valores que especificam o grau de influência que cada nodo tem sobre os demais. A mera inspeção dessas informações não permite ao teórico identificar os algoritmos que emergiram a partir do treinamento. Esses algoritmos só se tornam disponíveis para análise teórica quando um sistema computacional opera com base no modelo neural em questão. Em outras palavras, o teórico precisa analisar o comportamento resultante. Assim, ainda que o resultado dessa análise seja uma especificação dos algoritmos que o sistema segue por se deixar guiar pelo modelo neural, fica claro que a relação entre esses algoritmos e os algoritmos de treinamento utilizados para gerar o modelo *não é de reelaboração*. O exercício de certo conjunto de habilidades H capacita o sistema a desenvolver um outro conjunto de habilidades T, mas T não pode ser decomposta em articulações das habilidades presentes em H. Isso porque a caracterização de T depende fundamentalmente das informações disponibilizadas para o treinamento. Diferentes *corpora* de treinamento resultam em diferentes conjuntos de habilidades T, mesmo quando as habilidades exercitadas no treinamento (H) são rigorosamente as mesmas. Segue-se disso que não há um **V2** que especifique os algoritmos comuns às práticas que caracterizam o aprendizado e às práticas que emergiram desse aprendizado. Temos assim um meio de abarcar a transição entre duas situações **P1** e **P1'** que resulta de um treinamento,

como quer Brandom, mas que, contra Brandom, faz sentido mesmo na ausência de um metavocabulário **V2**.

Dito de outro modo, se a solução para o *frame problem* vier de uma prática de treinamento, como sugere Brandom, e se essa prática for compreendida no modelo do ML, como parece razoável, não existe um metavocabulário pragmático capaz de especificar a transição entre práticas – há, portanto, uma ruptura no modelo mesmo de Brandom. No entanto, como veremos na próxima seção, a exigência de interpretabilidade da IA nos traz de volta ao projeto de Brandom em MIE, que, no entanto, não é compreendido, ou pelo menos não inteiramente compreendido, com a chave proposta em BDS. A reorganização da compreensão de Brandom talvez seja maior do que sugerido aqui, já que a IA aparece no projeto de Brandom apenas em BDS, e não em MIE.

6

A adoção de modelos neurais gera uma consequência adicional significativa para o projeto de Brandom. A mecânica da explicitação proposta em *BSD* não se aplica a modelos neurais: para tais modelos, não existe um vocabulário **V2** que especifique os algoritmos que resultam de um aprendizado por treinamento por *ML*. Como já observamos, para nosso argumento, é irrelevante se a solução brandomiana para o *frame problem* é viável ou não – não é preciso que sua proposta tenha um alcance geral, basta que ela se aplique a pelo menos uma classe de transições entre práticas **P1** e **P1'** para as quais não exista um metavocabulário **V2**, e que essas transições sejam explicadas por um processo de treinamento. Nesse caso, ainda que o modelo geral de *BSD* não se aplique, nós reencontramos a solução brandomiana parcial via treinamento. Apesar dessa inadequação do modelo de *BSD* aos casos de extensão de práticas através de *frames* distintos, a motivação central de *MIE*, isto é, a busca do controle racional das inferências subjacentes aos comportamentos de um sistema, permanece e talvez mesmo ganhe em importância,

quando processos decisórios e, mais geralmente, o espaço de compreensão de fenômenos se devem à IA.

Trata-se de um problema bem conhecido entre os que se dedicam à regulamentação e à ética do uso da IA. Na medida em que esses sistemas participam de processos decisórios (quem deve ser contratado, quem deve ter o pedido de crédito financeiro aprovado etc.), é fundamental entender os motivos que levaram um modelo a decidir de uma forma ou outra. O esforço de tornar os processos decisórios mais compreensíveis é conhecido como o problema da interpretabilidade. A busca da interpretabilidade é o projeto da Inteligência Artificial Explicável (XAI, para '*Explainable Artificial Intelligence*').[XIV] Diante da pergunta "por que meu currículo foi rejeitado?", esperamos ouvir respostas como "você precisa ser formado na área X, não na área Y", mas também esperamos saber se e quando um currículo foi rejeitado por critérios raciais ou de gênero. Como é bem conhecido, na medida em que um modelo é treinado no *corpus* de decisões anteriores, ele tende a reproduzir os vieses presentes nessas decisões, ainda que isso nem sempre se mostre de forma clara.[XV] As razões para tais escolhas expressam não apenas uma certa cadeia de inferências que calhou de resultar numa rejeição, mas também os pesos de diferentes fatores em decisões desse tipo. Trata-se de um problema tão relevante que ele se tornou uma exigência de uma recente regulação da União Europeia: o direito à explicação.[XVI]

Por mais razoável que seja a exigência da interpretabilidade do ponto de vista político e ético, ela não resolve o problema teórico de tornar os algoritmos inteligíveis. Diferentes modelos podem apresentar diferentes relações entre interpretabilidade e *performance*, a depender dos

XIV Autores como Miller (2019) entendem que os termos "interpretabilidade" e "explicabilidade" apontam para *desiderata* diferentes. Essa distinção está geralmente associada ao objetivo de formular uma definição matemática rigorosa que permita, por exemplo, mensurar o quão explicável ou interpretável é um dado modelo. Como essa discussão é tangencial aos propósitos desse artigo, utilizamos aqui os termos indistintamente.

XV Ver Christian, 2020, pp. 17-117.

XVI "*The GDPR's policy on the right of citizens to receive an explanation for algorithmic decisions highlights the pressing importance of human interpretability in algorithm design*" (Goodman; Flaxman, 2017, p. 50).

domínios envolvidos. Especificar qual a região de uma imagem que levou o sistema a classificar seu conteúdo, por exemplo, pode tornar o processo mais compreensível e passível de entrar no jogo de dar e receber razões. Contudo, a mesma técnica não é elucidativa no caso de sequências complexas de DNA/RNA. Além disto, a interpretabilidade é sempre relativa a um agente: tornar o mecanismo que levou a uma decisão compreensível a um especialista não é *ipso facto* torná-lo acessível ao público mais geral, nem, portanto, aberto à discussão mais ampla. Como pano de fundo desses dois problemas, resta ainda a dúvida acerca da viabilidade teórica de uma exigência desse tipo: não é certo que algoritmos sedimentados em modelos neurais possam ser completamente especificados linguisticamente ou por meio de modelos mais simples.[XVII] Mas para compreender esse desafio, devemos ver o que se pode esperar da XAI.

Segundo Phillips et al. (2021), os princípios da XAI são os seguintes:

> i) O sistema deve ter as evidências ou razões para seus *outputs* e processos;
> ii) O sistema deve fornecer informações que os usuários possam compreender;
> iii) A explicação deve refletir de maneira acurada o modo como sistema funciona e gera seus *outputs*;
> iv) O sistema deve mostrar seus limites de funcionamento.[XVIII]

Sob mais de um aspecto, a *XAI* não é tão distante do que se espera das práticas conversacionais que trazem ao controle racional os comprometimentos inferenciais implícitos em decisões humanas. O princípio (i) é a motivação central para o projeto de *MIE*. A aplicação dos outros

XVII Abrahão et al. (2021), por exemplo, sugerem haver casos em que certos algoritmos que podem ser melhor expressos em redes neurais multidimensionais não podem ser especificados adequadamente em redes neurais com menos dimensões.

XVIII (Phillips et al., 2021, pp. 2–3) "First, the characterization needs to be human-centered, because humans consume them. Second, they need to be understandable to people. Third, explanations should correctly reflect the system's process for generating the output. To foster confidence in explanations, the system should indicate when it is operating outside its designed conditions." (p. 1). Ver também Gunning et al. (2019) e Christian (2020).

princípios ao modo como humanos compreendem e explicam suas próprias decisões não é direta. Em relação a (iii), a justificação de comprometimentos de agentes humanos num dado contexto e o modo como eles geram uma determinada escolha é compatível com a relativa opacidade do funcionamento da mente. Os limites indicados em (iv) não precisam ser claros para o agente ele mesmo, em parte pelas mesmas razões apontadas em (iii), mas também porque as discussões frequentemente ocorrem entre agentes que se encontram num mesmo contexto, o que torna ao mesmo tempo difícil e inútil a delimitação dos limites desse contexto. Nos dois casos, no entanto, a saída é a mesma: numa conversa, os comprometimentos relevantes para a discussão são trazidos à tona sem que seja necessário compreender o funcionamento subjacente da mente de cada agente, assim como são tornados explícitos os limites contextuais de determinadas formas de pensar. De resto, é precisamente isto o que deveríamos esperar no quadro do *MIE*.

O ponto (ii), por sua vez, revela uma distância maior entre a discussão entre agentes humanos e a XAI. Quando agentes se engajam na discussão acerca de seus comprometimentos inferenciais, eles o fazem nos termos que eles mesmos compreendem. A acessibilidade dos algoritmos que são explicitados pela XAI, por sua vez, deve ser um parâmetro ele mesmo explicitamente considerado e não é de modo alguma garantida.[XIX]

A XAI constitui um domínio efervescente. Tentativas de sintetizar os avanços e desafios sob uma *framework* unificada emergem quase todo ano. Nesse cenário ainda desorganizado, boa parte dos pesquisadores supõe que modelos opacos têm desempenho superior, e por isso defendem que a XAI deve se concentrar em análises *post hoc* das performances obtidas. Não por acaso, há grande quantidade de abordagens desse tipo, mas até o momento nenhuma delas consegue eliminar o risco de explicações inconsistentes com o processo decisório em questão. Parte do problema é que as informações trazidas à tona em análises *post hoc* são frequentemente (e talvez irremediavelmente), vagas. No exemplo de Longo et al. (2023), quando uma análise sugere que os dentes são parte

XIX Ver, e.g., Gunning et al. (2019) e Poursabzi-Sangdeh et al. (2021).

do que o modelo considera relevante para classificar alguém como um jovem adulto, não sabemos *como* a informação presente nessa região da imagem está sendo usada. O que constitui a pista que o modelo segue é a cor dos dentes ou a forma de um sorriso?[XX] Por ora, não está claro se e como esse tipo de dificuldade pode ser plenamente superada.

Quaisquer que sejam as abordagens envolvidas, elas devem trazer à tona o objetivo central da explicitação de comprometimentos: tornar as razões de uma determinada decisão explícitas, de modo que os sujeitos envolvidos possam compreender o por quê de escolhas que os concernem e, eventualmente, torná-las objetos de uma discussão pública. O ajuste de foco permite uma interpretação mais precisa da exigência (iii) ('a explicação deve refletir de maneira acurada o modo como sistema funciona e gera seus *outputs*'). O que está em jogo não é exatamente a compreensão dos algoritmos utilizados numa dada situação, mas os fatores aos quais tais algoritmos são sensíveis. Se a exigência da acurácia permanece, ela não diz respeito exatamente à compreensão dos mecanismos gerados por ML. A tarefa contínua não sendo fácil, mas não é exatamente a mesma de uma suposta explicitação de algoritmos, como talvez se poderia esperar a partir do modelo de Brandom em *BSD*.

Uma vez que ajustamos um pouco o foco, vemos que as preocupações do projeto do MIE e da XAI não são tão distintas. De fato, talvez se trate em parte de uma diferença de escala – decisões gerais, para a IA, e trocas conversacionais, para *MIE*. Para utilizar um vocabulário comum, podemos dizer que a XAI busca trazer os compromissos incorporados nos diferentes processos guiados pela IA para a conversa mais ampla das sociedades.

Para essa conversa mais ampla, há duas diferenças em relação à *MIE*. Inicialmente, os pontos (iii) e (iv) devem estar sob um controle contínuo, não para a busca da (talvez impossível) transparência algorítmica,

XX Há quem critique este caminho e defenda que a única XAI possível se dá pela rejeição de modelos opacos e a adoção de modelos intrinsecamente explicáveis, argumentando que isso não implica desempenho inferior (Rudin, 2018). Contudo, perseguir esse caminho nos afastaria do escopo das redes neurais e da solução de Brandom para o *frame problem*.

mas em razão da independência da IA em relação à agência humana – a IA sempre pode responder a fatores que nos são inteiramente opacos. A segunda diferença é que sempre há uma mediação entre o que chega à discussão pública e o que é utilizado num determinado processo decisório. Trata-se da mediação de uma abordagem *post hoc* utilizada para ampliar a interpretabilidade de um modelo. A escolha dessa abordagem se deve a um conhecimento especializado. Esse é o caso mesmo na hipótese de, num dado domínio, os modelos utilizados responderem a fatores que são mais ou menos transparentes aos usuários de um sistema. Essa conversa mais ampla levanta duas questões suplementares. A primeira pergunta diz respeito à relação entre a eficácia de uma solução do ponto de vista computacional e o que é exigido em decisões públicas. É possível que processos subótimos satisfaçam mais a critérios de explicabilidade sem perdas significativas para o que é relevante para decisões no espaço público. A segunda questão diz respeito ao que esperar da interpretabilidade. O que se busca não é exatamente a transparência dos algoritmos, mas a especificação das *features* a que eles respondem, ainda que o modo como eles o fazem permaneça opaco. A opacidade dos processos é, de resto, um traço pervasivo da cultura humana, dada a divisão do trabalho cognitivo.[XXI]

Essas consequências se afastam do projeto de *MIE*, ainda que busquem o mesmo objetivo, com o ajuste de escala sugerido a acima. Elas também mostram que se o modelo de *BSD*, como já vimos, não se aplica a redes neurais que se desenvolvem por *ML*, ele coloca uma pergunta crucial, a saber, qual o algoritmo ou, de maneira talvez mais neutra, quais os comprometimentos em jogo em cada situação trazida à baila por uma discussão e o que, desses algoritmos, pode ser tornado explícito. A resposta de *BSD* não está certa nem para a agência humana, nem para a IA, mas talvez o modelo de Brandom traga boas perguntas, senão para a agência humana, pelo menos para a IA propriamente dita.

XXI Conforme Perini-Santos (2022; 2025).

7

A associação entre a abordagem de Brandom para mitigar os efeitos do *frame problem* e redes neurais é intrigante e traz consequências importantes para a compreensão do seu projeto. A primeira delas é que não estamos mais no quadro da IA clássica. A segunda consequência é que a natureza de **V2** muda: não se trata mais da especificação de algoritmos utilizados na execução de uma tarefa, mas de algoritmos que especificam o treinamento por meio do qual outras capacidades algorítmicas emergem, sem que a relação entre estes seja caracterizada como de re-elaboração. Talvez se possa dizer que se trata de um vocabulário **V2** com um alcance mais restrito, mas isto não capta a extensão do problema - sem saber a evolução do treinamento de algoritmos, não há como afirmar que o resulto obtido tenha um papel fundacional para a compreensão do modo como agimos e falamos. [XXII]

Efetivamente, nesse cenário, a preocupação com abordagens que respeitem características biológicas ou psicológicas torna-se ainda mais aguda. Não basta que o modelo treinado expresse um modo caracteristicamente humano de exibir uma dada habilidade ou performance. É preciso que o processo de treinamento do modelo atenda os mesmos requisitos. Há, sabidamente, diferenças importantes entre o modo como seres humanos aprendem e a forma como algoritmos de aprendizado atuam (Dehaene, 2021), e a mitigação destas diferenças não é uma preocupação típica na IA contemporânea. [XXIII] Isso significa que o uso de um dado algoritmo de ML precisa ser cuidadosamente avaliado caso a caso. Se não houver preocupação em replicar as condições de aprendizagem que caracterizariam o desenvolvimento de nossos mecanismos cognitivos, o que perdemos aqui é o interesse mesmo da solução sugerida por Brandom para uma análise filosófica.

XXII De fato, a possibilidade de compreender como uma determinada capacidade emerge a partir do ML é, ela mesma, objeto de pesquisa dentro da IA. Ver por exemplo Nanda et al. (2023).

XXIII Para exemplos de esforços que podem ser úteis a esse fim, ver Stinson (2018) e Buckner (2023).

A terceira consequência é que, no caso de modelos neurais, pode não haver nada que se assemelhe a uma prática de explicitação nos termos requeridos pelo projeto de *BSD*. Com efeito, ainda que tais modelos capturem um maior número de casos, não é claro se a opacidade do funcionamento desses modelos pode ser superada de um modo frutífero para Brandom, pois as ocasiões em que a explicitação parece necessária (do ponto de vista do *MIE*), continuam descobertas.

As duas sugestões que vimos anteriormente para a extensão da tese de Hurley oferecem dois modelos pelo menos parciais para as *constraints* impostas sobre a IA como modelagem da capacidade raciocínio humano.

Uma primeira resposta é que os limites de soluções algorítmicas exprimíveis na forma de conjuntos de regras refletem os limites humanos de pensar através de situações diferentes. O problema é que mesmo se admitirmos que o raciocínio humano não é tão flexível quando poderíamos esperar de um agente capaz de integrar todas as suas crenças, ele é certamente capaz de resolver muitas transições entre *frames* que geram problemas computacionalmente intratáveis. A ameaça da intratabilidade não se coloca apenas em sistemas capazes de integrar crenças de forma irrestrita. Basta que o grau de integração e flexibilidade disponíveis resulte num conjunto de inferências grande demais para que a consideração exaustiva de todas as possibilidades seja computacionalmente plausível. Isso já obriga o sistema a determinar, de algum modo, quais dentre todos os fatores possíveis são circunstancialmente relevantes quando considerando uma transição entre *frames*, e essa capacidade supõe uma solução para o *frame problem*. Não é preciso que o mundo inteiro constitua um sistema isotrópico.

A segunda resposta é que há uma heterogeneidade no modo como pensamos sobre diferentes situações e talvez funcionemos, em muitos contextos, de uma maneira mais próxima do que, nos modelos neurais, permanece opaco. Como discutido na seção anterior, tanto os feitos quanto os limites da IA contemporânea podem ser interpretados como um sinal de que nossas práticas de explicitação têm uma natureza distinta da que Brandom supõe.

Nenhuma dessas opções parece poder ser acomodada no projeto de Brandom, pelo menos não na versão proposta em *BSD*. Ambas as redescrições, contudo, parecem reservar um lugar diferente para a explicitação daquele que encontramos em *BSD*, mas, talvez de maneira inesperada, próximo do modelo proposto em *MIE*. O projeto da *XAI* de tornar mais interpretáveis os caminhos que levam a determinadas decisões não visam exatamente tornar os algoritmos mais transparentes. De fato, se, à maneira do vocabulário metapragmático de *BSD*, pensamos na descrição dos algoritmos utilizados num mecanismo de decisão, a tarefa é simplesmente impossível para os algoritmos emergentes pelo aprendizado de máquina (e, lembremos, o aprendizado de máquina parece ser a resposta compatível com a solução de Brandom para o *frame problem*). No entanto, se o que buscamos é uma maneira de trazer nossas práticas para o controle racional, "*in a form in which they can be confronted with objections and alternatives*" (*MIE*, 106), basta um passo suplementar na descrição dos algoritmos, mas não sua especificação completa – em outros termos, aumentar a interpretabilidade de processos de tomada de decisão não é especificar os algoritmos utilizados, mas tornar salientes os passos que levam a caminhos que podem ser corrigidos, por exemplo, à luz de considerações políticas e morais.

8

O papel fundacional da IA na filosofia de Brandom surge na confluência de alguns temas que lhe são caros: a teoria pragmatista do significado, a força teórica do projeto da semântica formal, mas também a explicitação da rede de comprometimentos que, numa semântica inferencialista, fornece o significado de nossos diferentes vocabulários. Os algoritmos que descrevem os comportamentos que constituem o domínio de um vocabulário são precisamente a realização, por um programa de computador, do exercício da capacidade de pensar sobre um determinado tema, isto é, o programa da IA clássica (*BSD*, 70).

Colocado nos termos de BSD, os algoritmos constitutivos de uma prática **P1** que fundamenta um vocabulário **V1** são especificados pelo vocabulário metapragmático **V2**. O vocabulário **V2** ele mesmo é fundamentado numa prática **P2** que é a reorganização dos algoritmos que constituem **P1** para uma outra função. **P2** não dá aos agentes as mesmas capacidades exercidas em **P1**, mas fornece um ponto de vista metateórico sobre essas mesmas práticas, que permite não apenas explicitar os direitos e comprometimentos que estruturam **P1**, mas também tornar esses direitos e comprometimentos o objeto de discussão. Aqui está a conexão dos projetos de BSD e MIE: a explicitação dos comprometimentos inferenciais que surgem no jogo de dar e receber razões parece não ser outra coisa senão **P2**.

O projeto de Brandom, no entanto, falha por uma razão já bem conhecida na IA: não é possível dar uma descrição algorítmica das práticas humanas, em razão de nossa capacidade de transitar entre diferentes situações que exigem o fino ajuste de direitos e comprometimentos, o que é conhecido como *frame problem*. De fato, se formos especificar as regras que guiam nosso comportamento através de variações de situações que nos parecem triviais, rapidamente nos encontramos diante de uma explosão combinatorial que torna impossível a consideração exaustiva de todas as possíveis articulações inferenciais. Colocado nos termos de BSD, não existe um metavocabulário **V2** que explique a transição entre *frames* que mobilizam exercícios inferenciais distintos.

Brandom propõe uma solução para o *frame problem* através de aprendizado por treinamento. De maneira crucial, o treinamento que permite a transição entre capacidades inferenciais não é guiado por algoritmos especificados num metavocabulário **V2**. Essa sugestão sai do quadro geral proposto em BSD. Não se trata de uma consequência menor para a articulação de MIE e BSD: o interesse de tornar explícitos comprometimentos inferenciais surge, precisamente, em casos de ruptura de um determinado tipo de comportamento. Ora, essas são as situações nas quais a solução de Brandom não funciona, ao exigir a especificação de algoritmos que constituem as práticas humanas.

Há, no entanto, uma outra leitura para o projeto, ou para os projetos de Brandom. De fato, podemos ler a transição entre *frames* através de treinamento como a extensão de redes neurais através de ML. Essa exige a rearticulação da relação entre a filosofia de Brandom e a IA – não estamos mais no quadro da IA clássica e não é tão claro em que sentido processos que são pelo menos parcialmente opacos podem servir como fundamentação de práticas humanas. No entanto, a exigência de explicitação de MIE se mantém, e é mesmo reforçada: na mesma medida em que aumenta o espaço ocupado pela IA na organização do nosso mundo comum, também cresce a importância de tornar explícitos comprometimentos inferenciais, ao mesmo tempo onipresentes e quase totalmente opacos para quase todo mundo. A mecânica da explicitação da MIE aplicada à IA assume a forma da XAI, não mais como a especificação de algoritmos que constituem práticas humanas, mas como a tarefa suplementar de torna legível a articulação do espaço inferencial da IA. Se as respostas de Brandom na confluência entre MIE e BSD não parecem suficientes, as perguntas que ele traz são cada vez mais prementes.).

REFEFRÊNCIAS

ABRAHÃO, F. S. et al. An algorithmic information distortion in multidimensional networks. In: *Studies in Computational Intelligence*. Cham: Springer International Publishing, 2021. p. 520–531.

ARJOVSKY, M. Out of distribution generalization in machine learning. *arXiv preprint* arXiv:2103.02667, 2021.

BARTH, C. *Representational cognitive pluralism*: *towards a cognitive-science of relevance-sensitivity*. Tese (Doutorado) — Belo Horizonte: Faculdade de Filosofia e Ciências Humanas, Universidade Federal de Minas Gerais, 2024.

BARTH, C. *O "frame problem"*: *a sensibilidade ao contexto como um desafio para teorias representacionais da mente*. Dissertação (Mestrado) — Belo Horizonte: Faculdade de Filosofia e Ciências Humanas, Universidade Federal de Minas Gerais, 2019.

BARTH, C. É possível evitar vieses algorítmicos? *Revista de Filosofia Moderna e Contemporânea*, v. 8, n. 3, p. 39–68, 2021.

BRANDOM, R. B. *Between saying and doing*: *towards an analytic pragmatism*. Oxford: Oxford University Press, 2008.

BRANDOM, R. B. *Making it explicit*: *reasoning, representing and discursive commitment*. Cambridge, MA: Harvard University Press, 1994.

BUCKNER, C. J. *From deep learning to rational machines*: *what the history of philosophy can teach us about the future of artificial intelligence*. Oxford: Oxford University Press, 2023.

CHRISTIAN, B. *The alignment problem*. New York: W. W. Norton & Company, 2020.

CHURCHLAND, P. S. *Neurophilosophy*: *toward a unified science of the mind-brain*. Cambridge, MA: MIT Press, 1989.

CLARK, A. Local associations and global reason: Fodor's frame problem and second-order search. *Cognitive Science Quarterly*, n. 2, p. 115–140, 2002.

CUPANI, A. *Filosofia da tecnologia*. Florianópolis: Editora UFSC, 2011.

DEHAENE, S. *How we learn*. London: Penguin Books, 2021.

DENNETT, D. Cognitive wheels: the frame problem of AI. In: PYLYSHYN, Zenon W. (ed.). *The robot's dilemma*: *the frame problem in artificial intelligence*. Norwood, NJ: Ablex, 1987. p. 41–64.

DREYFUS, H. *What computers still can't do*. Cambridge, MA: MIT Press, 1992.

GEIRHOS, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint* arXiv:1811.12231, 2019.

GOODMAN, B.; FLAXMAN, S. European Union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, v. 38, n. 3, p. 50–57, 2017.

GUNNING, D. et al. XAI—Explainable artificial intelligence. *Science Robotics*, v. 4, n. 37, 2019.

HASELAGER, W. F. G.; RAPPARD, J. F. H. V. Connectionism, systematicity, and the frame problem. *Minds and Machines*, v. 8, p. 161–179, 1998.

HURLEY, S. Making sense of animals. In: HURLEY, Susan; NUDDS, Matthew (eds.). *Rational animals*? Oxford: Oxford University Press, 2006. p. 139–171.

LIU, J. et al. Towards out-of-distribution generalization: a survey. *arXiv preprint* arXiv:2108.13624, 2021.

LONGO, L. et al. Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. *arXiv preprint* arXiv:2310.19775, 2023.

MCCARTHY, J.; HAYES, P. J. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, v. 4, p. 463–502, 1969.

MILLER, T. Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence*, v. 267, p. 1–38, 2019.

MINSKY, M. A framework for representing knowledge. In: HAUGELAND, John (ed.). *Mind design II*: *philosophy*, *psychology*, *artificial intelligence*. Cambridge, MA: MIT Press, 1997. p. 111–142.

NANDA, N. et al. Progress measures for grokking via mechanistic interpretability. *arXiv preprint* arXiv:2301.05217, 2023.

PERINI-SANTOS, E. Desinformação, negacionismo e a pandemia. *Filosofia Unisinos*, v. 23, n. 1, p. 1–15, 2022.

PERINI-SANTOS, E. *Viver é fácil de olhos fechados*: *fake news*, *negacionismo e teorias da conspiração*. Belo Horizonte: Editora UFMG, 2025.

PHILLIPS, P. J. et al. Four principles of explainable artificial intelligence. *National Institute of Standards and Technology* (*U.S.*), 2021.

PICCININI, G. *Physical computation*: *a mechanistic account*. Oxford: Oxford University Press, 2015.

POURSABZI-SANGDEH, F. et al. Manipulating and measuring model interpretability. In: *Proceedings of the* 2021 *CHI Conference on Human Factors in Computing Systems* (*CHI '*21). New York: ACM, 2021.

RUDIN, C. Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *arXiv preprint* arXiv:1811.10154, 2018.

RYLE, G. 'If', 'So', and 'Because'. In: BLACK, Max (ed.). *Philosophical analysis*. Ithaca, NY: Cornell University Press, 1950. p. 323–340.

SAMUELS, R. Classical computational models. In: SPREVAK, Mark; COLOMBO, Matteo (eds.). *The Routledge handbook of the computational mind*. London: Taylor & Francis, 2018. p. 103–119.

SCHANK, R. C.; ABELSON, R. P. *Scripts*, *plans*, *goals and understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1977.

SCHARP, K. A. Scorekeeping in a defective language game. *Pragmatics & Cognition*, v. 13, n. 1, p. 203–226, 2005.

SHANAHAN, M. *Solving the frame problem*: *a mathematical investigation of the common sense law of inertia*. Cambridge, MA: MIT Press, 1997.

SMITH, B. C. *The promise of artificial intelligence*: *reckoning and judgment*. Cambridge, MA; London: MIT Press, 2019.

STINSON, C. Explanation and connectionist models. In: SPREVAK, Mark; COLOMBO, Matteo (eds.). *The Routledge handbook of the computational mind*. London: Routledge, 2018.

Recebido em 01 de agosto de 2024
Aprovado em 15 de junho de 2025
Publicado em 21 de outubro de 2025