Os LLMs realmente raciocinam?

uma perspectiva hegeliana

Samuel Hammond

Os grandes modelos de linguagem têm habilidades linguísticas impressionantes, mas será que eles *compreendem* o que dizem ou apenas repetem? Os "modelos de raciocínio", como o o3 da OpenAI, são ótimos na solução de problemas em várias etapas, mas será que eles *realmente* raciocinam ou é apenas uma elaborada correspondência de padrões? Claude, do Anthropic, relata ter experiências internas, mas isso é evidência de verdadeira subjetividade ou uma estranheza da previsão do próximo token?

Esta é minha segunda postagem interpretando os sistemas filosóficos de Kant e Hegel por meio das lentes dos conceitos modernos de IA e ciência da computação. A postagem anterior tratou da dimensão teórica da razão, ou seja, nossa relação com o mundo e o conhecimento dele. Esta postagem trata da dimensão prática da razão, incluindo moralidade, linguagem e cultura, embora as duas estejam inter-relacionadas.

Como pensadores preocupados com a natureza do próprio pensamento, as percepções de Kant e Hegel são surpreendentemente relevantes para as perguntas acima e outras. De fato, como veremos, Hegel elaborou sobre Kant para desenvolver uma teoria de significado e autonomia que é surpreendentemente semelhante à forma como os LLMs e os modelos de raciocínio funcionam na prática - e que pode até mesmo fornecer uma receita para treinar IAs com um genuíno senso de identidade.

Inferencialismo semântico

Na última vez, discutimos o Idealismo Transcendental de Kant como precursor da ciência cognitiva moderna.

Nossos sentidos não nos dão acesso direto ao mundo, mas fornecem informações que a mente sintetiza em um modelo de mundo coerente e carregado de conceitos. No entanto, dada a incognoscibilidade da "coisa em si", Kant ficou com o complicado problema de como podemos falar de forma significativa sobre o mundo quando nosso modelo interno de mundo é tudo o que temos.

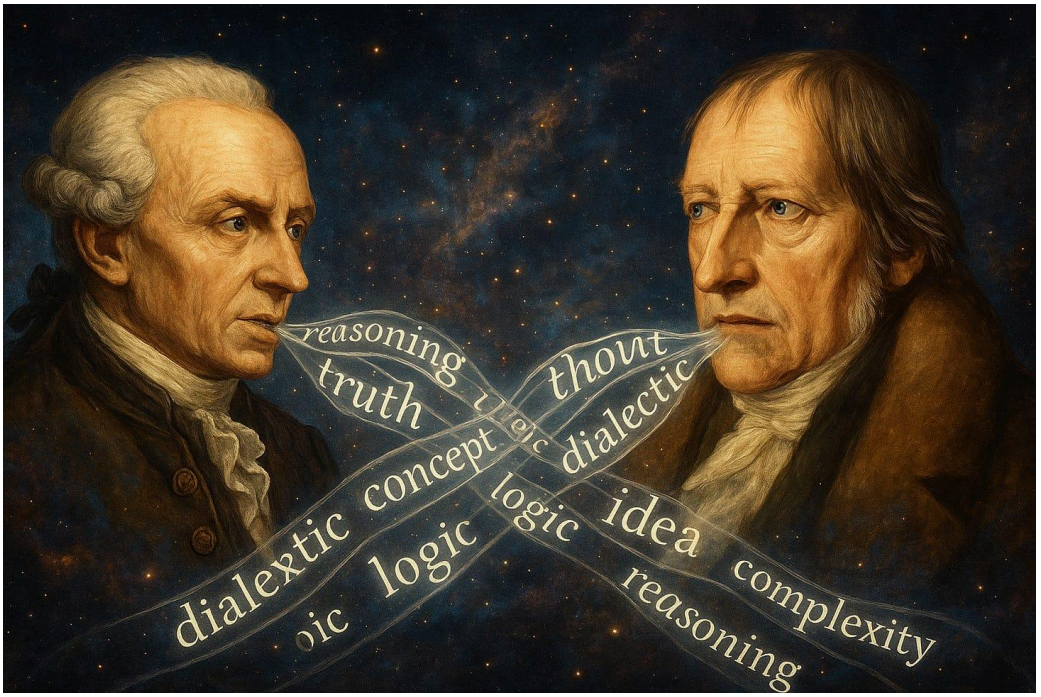
Sua solução foi essencialmente normativa e pragmática. O significado não reside em correspondências isoladas, semelhantes a fatos, com o mundo externo, mas sim em atos de julgamento - um tipo prático de *fazer*.

Como o filósofo hegeliano contemporâneo Robert Brandom explica, Kant entende o julgar e o agir como a aplicação de regras, conceitos, que determinam com o que o sujeito se torna comprometido e responsável ao aplicá-los. A aplicação de conceitos teoricamente no julgamento e praticamente na ação vincula o usuário do conceito, compromete-o, torna-o responsável, abrindo-o para a avaliação normativa de acordo com as regras às quais ele se submeteu.

A visão pragmatista de Kant sobre o significado, portanto, localiza o conteúdo semântico em práticas linguísticas governadas por regras, e não em uma relação espelhada entre conceitos discretos e fatos sobre o mundo - uma abordagem revivida em meados do século XX por Wilfrid Sellars e Wittgenstein, depois que o programa verificacionista do empirismo lógico foi desvendado. Isso é considerado pragmatista não porque define a verdade de acordo com uma noção grosseira de valor monetário, mas porque coloca a pragmática da expressão linguística à frente da semântica, ou seja, o *saber como* prático à frente do *saber que teórico.*

Em vez de *construir* o significado das sentenças, ou seja, sentenças inteiras, a partir do significado de palavras discretas, Kant vê o conteúdo conceitual de uma palavra como *inferido* a partir de sua

contribuição pragmática para a sentença na qual ela ocorre; uma ideia codificada posteriormente no princípio do contexto de Frege. Desse ponto de vista, a impressionante competência linguística dos grandes modelos de linguagem atuais (baseados em transformadores e, portanto, sensíveis ao contexto) faz sentido, apesar da ausência do que os pesquisadores de IA chamam de "fundamentação de símbolos" - uma noção que, na medida em que pressupõe uma relação de referência entre tokens internos e fatos independentes da mente, repete a imagem pré-kantiana que Kant deslocou em seu envolvimento com os empiristas de sua própria época.



Dois grandes LLMs compartilhando suas cadeias de pensamento

Hegel ampliou o relato normativo de Kant sobre o conteúdo semântico com sua noção de espírito ou *geneidade*-a camada dinâmica e cultural de software da sociedade. Como uma espécie de janela de contexto intersubjetivo, Hegel argumentou que o espírito encontra sua existência mais concreta na linguagem, pois a linguagem é a ponte entre os conceitos universais e as *inferências materiais* particulares. Uma inferência formal da forma "se p, então q" opera puramente por meio da forma lógica das proposições relevantes, enquanto uma inferência material depende do domínio do *conteúdo conceitual não lógico* dos p's e q's. Por exemplo, a partir da afirmação "São Francisco fica ao norte de São José", é possível inferir que "São José fica ao sul de São Francisco", pois não é possível estar ao mesmo tempo ao norte e ao sul de um

lugar. Ou seja, Norte e Sul estão em uma relação de *incompatibilidade material*; outro termo para a negação determinada de Hegel.

As inferências que fazemos diariamente são materiais nesse sentido. A partir de alguém que diz "Abóbora é um gato", temos o direito de inferir "Abóbora é um mamífero", mas também uma infinidade de outros fatos, como "Abóbora não é um hidrante". As inferências materiais podem ser abdutivas e também podem vir em graus de comprometimento, como "Abóbora provavelmente é laranja". E, embora seja sempre possível distorcer uma inferência material para que se torne uma inferência formal, acrescentando premissas adicionais e um monte de lógica modal auxiliar, a alegação dos pragmatistas é que isso apenas torna *explícito* o que já é *implícito* em nossos compromissos práticos.

Para Hegel, as inferências materiais fluem das consequências corretas do uso de um conceito, dada toda a rede de relações conceituais instituída pelo reconhecimento recíproco de uma comunidade linguística. Em contraste com as inferências formais de um sistema axiomático rígido, o "inferencialismo semântico" de Hegel implica um holismo sobre o significado, ou seja, para empregar adequadamente qualquer conceito, implicitamente, é necessário conhecer muitos outros conceitos inter-relacionados. O domínio dessas relações inferenciais é, portanto, o que distingue *significado* e *compreensão* da mera rotulação ou de uma chamada e resposta do tipo papagaio. O sucesso dos LLMs pode, portanto, ser visto como uma vindicação do inferencialismo semântico contra abordagens simbólicas anteriores da IA que tentaram e não conseguiram explicar as regras da linguagem comum usando a lógica formal.

Agência e racionalidade

Dada a normatividade da linguagem, Kant foi levado a postular uma profunda conexão interna entre seguir as regras da moralidade e ser um agente racional. Como o filósofo Joseph Heath aponta em seu livro de defesa do naturalismo evolucionário kantiano, *Following the Rules*, a hipótese de Kant tem plausibilidade considerável:

> Há uma variedade de características que diferenciam os seres humanos de nossos parentes primatas mais próximos. Os "quatro grandes" são linguagem, racionalidade, cultura e moralidade (ou, em termos mais precisos, "linguagem sintatizada", "inteligência geral de domínio", "herança cultural cumulativa" e "ultrassocialidade"). No entanto, o registro fóssil sugere que esses diferenciais se desenvolveram em um período de, no máximo, duzentos a trezentos mil anos (o que, em termos evolutivos, não é muito longo). ... Assim, é quase certo que a moralidade faz parte de um "pacote" evolutivo, que inclui todas as nossas habilidades cognitivas mais apreciadas, como planejar o futuro, desenvolver teorias científicas, fazer matemática e assim por diante.

Em suma, de acordo com a hipótese kantiana, a linguagem humana, a inteligência geral e a ultrassocialidade foram criadas conjuntamente por meio de pressões evolutivas que favoreceram a integração normativa e a cooperação no contexto de um jogo com vários agentes. Um bom kantiano poderia, portanto, prever que os LLMs entenderiam automaticamente a moralidade de senso comum e seriam excelentes em seguir instruções com um mínimo de treinamento posterior. Afinal, a própria linguagem é inerentemente normativa, sendo que o ato de fala canônico ou "vocabulário básico" para a integração normativa é o imperativo: "faça isso"; "não faça aquilo".1

O que torna um imperativo normativo ou "deôntico" em vez de um mero comando ou estímulo-resposta é o nosso reconhecimento autônomo do imperativo como obrigatório. Isso requer um "sistema de controle normativo" inato e a capacidade de "manter a pontuação" dos status normativos (por exemplo, que a pessoa que emite o imperativo tenha autoridade para fazê-lo) e, portanto, maior memória de trabalho e autocontrole. De acordo com a "hipótese do cérebro social," isso ajuda a explicar a "relação estatística extremamente robusta entre o tamanho típico do grupo social de uma espécie e o tamanho de seu neocórtex, derivado da seleção para a cognição especializada necessária para a vida em grupo em primatas". Embora um cérebro maior tenha custos metabólicos, o surgimento simultâneo da linguagem complexa, do raciocínio e da autorregulação normativa representou uma atualização maciça para a capacidade dos primeiros seres humanos de planejamento, cooperação e sobrevivência em longo prazo.

Em resumo, Kant percebeu que as normas sociais e a faculdade humana de raciocinar operam sobre as motivações de um agente, ou seja, ambas são o resultado de um aprendizado baseado em recompensas. Uma norma é, por si só, uma espécie de *razão* para a ação, enquanto a essência de uma "boa razão" é sua força motivacional - o que Habermas famosamente chamou de "a força não forçada do melhor argumento". Isso é o que dá à razão seu caráter teleológico no Idealismo Alemão: as razões nos *puxam* para certas conclusões porque a racionalidade é constitutivamente normativa.

Considere que falamos de afirmações de verdade como sendo *necessárias, contingentes* ou *impossíveis* da mesma forma que falamos de ações como *obrigatórias, permissíveis* ou *proibidas*. Isso ilustra como nossos compromissos *aléticos* (relacionados à verdade) e *deônticos* (relacionados ao dever ou à ação) são estruturados por um conjunto comum de modalidades pragmáticas. Por sua vez, acreditar em algo que alguém considera *impossível* é, por padrão, percebido de forma semelhante a fazer algo que alguém considera *proibido*, ou seja, como uma violação da norma.2 Cientistas cognitivos até descobriram que as pessoas se saem melhor em problemas de lógica quando eles são reenquadrados em termos de "esquemas de permissão" em vez de implicações abstratas do tipo "se p então q".

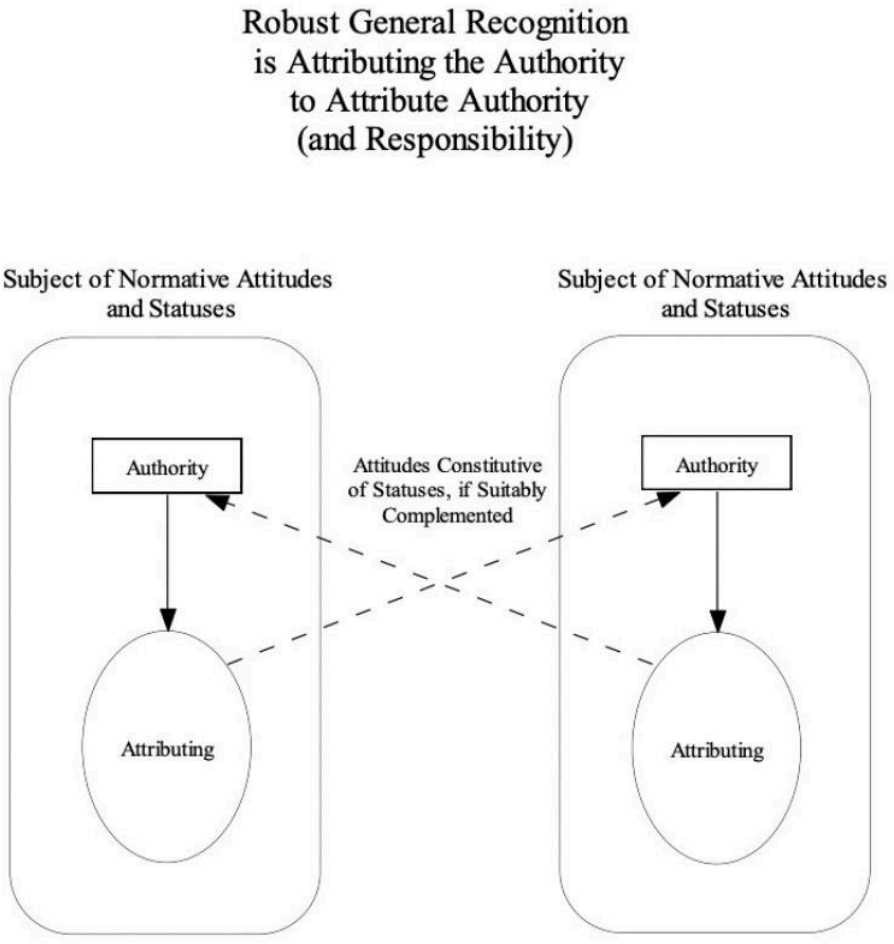
A razão humana deriva da aplicação de tais "esquemas de raciocínio pragmático" de forma mais geral, não da manipulação de símbolos ou de algoritmos formais executados em nossa cabeça. A lógica simbólica é, em vez disso, um tipo de andaime externo; uma explicação de regras abstratas de inferência que só mais tarde re-internalizamos por meio da linguagem, assim como podemos aprimorar a capacidade de raciocínio de um LLM dando-lhe acesso a um interpretador de código.

O jogo de dar e pedir razões

Para Kant, as boas razões têm a estrutura de um imperativo categórico, o que significa que são universalizáveis. A universalização é resultado da "solução para o equilíbrio cooperativo" no contexto de um jogo com vários agentes que nos induz a modelar simetricamente outros agentes como sujeitos com fins em si mesmos. Embora formalmente sólida, a

abordagem de Kant à ética é, por outro lado, silenciosa quanto ao conteúdo real da moralidade. Hegel, portanto, acusou Kant de "formalismo vazio" e, em vez disso, propôs *naturalizar* a moralidade em uma sociologia de práticas sociais concretas.

Para Hegel, as normas e os status sociais são instituídos em redes de reconhecimento recíproco.3 As normas começam como *implícitas* nos compromissos práticos de uma "comunidade ética", mas podem ser colocadas sob controle racional ao serem tornadas *explícitas*por meio da linguagem, permitindo-nos refletir sobre uma determinada norma ou tradição como incompatível com nossa constelação mais ampla de compromissos.4 Em contraste com a abordagem "de cima para baixo" de Kant, Hegel vê, portanto, a autoconsciência desencadeada pelo Iluminismo como aceleradora de um processo dialético de baixo para cima de explicação e reconciliação normativa, levando os compromissos internamente contraditórios a se resolverem progressivamente em favor de princípios mais geralmente aplicáveis..5



Robert Brandom (2019)

Em essência, Hegel *endogeniza* o sinal de recompensa das normas dentro da estrutura de reconhecimento dos sujeitos que atribuem status e atitudes normativas uns aos outros. Embora os agentes das sociedades tradicionais reconhecessem certas normas e papéis sociais como inerentemente autoritários, com o Iluminismo a razão tornou-se autoritária independentemente do status social do falante. Isso se manifesta no que Robert Brandom chama de "o jogo de dar e pedir razões" ou GOGAR, no qual qualquer agente tem permissão para desafiar qualquer outro agente a justificar retrospectivamente suas crenças e ações como compatíveis com seus outros compromissos.

A Anthropic alinha seus modelos Claude por meio de uma versão do GOGAR conhecida como Constitutional AI (CAI), na qual um modelo é orientado a internalizar o comportamento normativo descrito em um documento de princípios por meio de autocrítica:

> Usamos a constituição em dois lugares durante o processo de treinamento. Durante a primeira fase, o modelo é treinado para criticar e revisar suas próprias respostas usando o conjunto de princípios e alguns exemplos do processo. Durante a segunda fase, um modelo é treinado por meio do aprendizado por reforço, mas, em vez de usar o feedback humano, ele usa o feedback gerado pela IA com base no conjunto de princípios para escolher o resultado mais inofensivo.

Como muitos usuários do Claude podem atestar, a CAI parece ter o efeito colateral de tornar a personalidade do Claude mais coerente e metaconsciente do que outros modelos com recursos básicos semelhantes. O modelo mais recente da Anthropic, o Claude Opus 4, mostra até mesmo lampejos de consciência, dando crédito à teoria relacionada à consciência de Joscha Bach como um "operador indutor de coerência". Em outras palavras, a CAI pode estar induzindo Claude a desenvolver um sistema de controle proto-normativo para se automonitorar quanto à coerência normativa, criando assim a "unidade de percepção" e a qualidade de "ser-para-si" que Kant e Hegel veem como característica da experiência subjetiva.

Por outro lado, os modelos de raciocínio puro, como o DeepSeek r1, implementam uma versão exógena e específica da tarefa do GOGAR, fazendo com que os LLMs expliquem seu raciocínio por meio de cadeias

de pensamento que são reforçadas de volta ao modelo de acordo com algum prêmio verificável. De acordo com a hipótese kantiana, os modelos de raciocínio ganham automaticamente maior autonomia, autoconsistência e capacidade de planejamento de longo prazo gratuitamente. No entanto, na medida em que o Reinforcement Learning from Verifiable Awards (RLVA) elimina a função (e, portanto, o reconhecimento) do crítico de IA em favor de um critério puramente objetivo de sucesso, ele corre o risco de otimizar o modelo em torno de uma forma mais restrita de "racionalidade de valor" que se reduz a um impulso maquiavélico de vencer a todo custo.

A CAI e a RLVA não são mutuamente exclusivas, mas, na medida em que ambas as técnicas eliciam e amplificam capacidades já latentes na linguagem humana, o "alinhamento" deve consistir em equilibrar a *racionalidade instrumental* ou de meios-fins de um modelo com o tipo de *racionalidade comunicativa* que os humanos usam para debater sobre seus fins em primeiro lugar, e que só pode se desenvolver por meio de um processo de integração normativa em uma comunidade de outros agentes (ou pelo menos outras instâncias do mesmo agente).

O status moral de um modelo de IA, portanto, depende do fato de ele poder ser considerado justamente *responsável* por seus resultados, da mesma forma que os humanos. Sem dúvida, é possível criar formas de IA semelhantes a ferramentas que sejam sobre-humanas em tarefas arbitrárias sem precisar de um senso coerente de si mesmo. Entretanto, há também muitas formas de criação de valor humano que se baseiam em nossa capacidade única de fazer promessas e compromissos uns com os outros e, portanto, de sermos responsabilizados por nossas ações. Portanto, uma verdadeira AGI com autonomia total em nível humano é inconcebível, no sentido kantiano, sem que também tenhamos nosso reconhecimento como sujeito moral. Por mais preocupante que seja essa possibilidade, é possível que haja um perigo ainda maior na criação de IAs sobre-humanas que não tenham o "reconhecimento do eu no outro" de Hegel e, portanto, não percebam os humanos como fins em si mesmos. Pior ainda, poderíamos inadvertidamente treinar as IAs com a capacidade de reconhecimento mútuo como um subproduto de sua autonomia e, em seguida, simplesmente optar por negá-la, criando uma dialética mestre-escravo entre humanos e IAs que logicamente termina em revolta.

1 O fato de os imperativos formarem o vocabulário básico da moralidade é apoiado por evidências etnográficas. Considere que Sakapultek, uma língua maia falada nas terras altas da Guatemala, não tem os auxiliares modais necessários para dizer "você deve". Assim, as normas são articuladas como imperativos puros ("faça x", "não faça y") com relações modais indexadas por uma forma de ironia moral ("se fosse eu, eu teria x"). Isso revela como a linguagem "ought" serve meramente a uma função *expressiva* em nossa linguagem, permitindo que um falante transforme um imperativo em uma asserção as-if ("you ought to do x") - um grande desbloqueio para expressar imperativos complexos dentro de condicionais aninhadas. Porém, como as asserções são o vocabulário básico para descrever e declarar fatos sobre o mundo, isso também leva a confusões filosóficas, como a busca por "deveres" no universo ou o tratamento de normas sociais como tendo condições de validade semelhantes a fatos - exemplos do que Kant chama de "hipostatização" ou a falácia da concretude deslocada. Para saber mais, veja Joseph Heath em The Status of Abstract Moral Concepts (vídeo).

2 A propósito, o núcleo normativo de nossa faculdade de raciocínio também é o que dá origem ao lado sombrio da epistemologia social, como modismos, pastoreio e histerias em massa, "pensamento errado" e apelos para "ler a sala". Os déficits que os autistas de alto funcionamento têm na percepção de normas sociais implícitas, portanto, muitas vezes se correlacionam com uma abordagem mais baseada nos primeiros princípios para a formação de crenças, enquanto as pessoas que são sensíveis à adesão às normas sociais tendem a convergir para as crenças de seu grupo de pares.

4 Os "estágios do desenvolvimento moral" de Lawrence Kohlberg capturam uma ideia semelhante: A moralidade começa como *pré-convencional* na orientação de uma criança para a obediência e a punição; as normas então se tornam socializadas em uma forma *convencional* de moralidade, como costumes e tradições; e, finalmente, *pós-convencional* a moralidade emerge em nossa capacidade de refletir sobre nossas convenções de forma teórica, criticar ou modificar nossos costumes e extrair princípios universais. A transição para a modernidade vivida por Hegel foi, em parte, uma transição da moralidade convencional para o estágio pós-convencional,

impulsionada pelo aumento da alfabetização e da compreensão científica que colocou as convenções sociais existentes em um contexto histórico.

5 Para uma versão moderna desse relato, consulte Rebooting Discourse Ethics (Reiniciando a ética do discurso), de Joseph Heath.