Hegel e a mente da IA

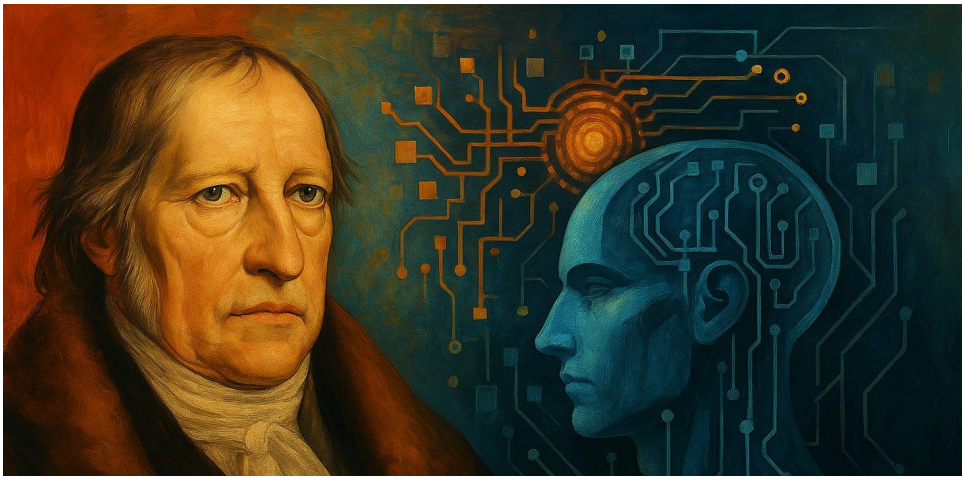
Reconstruindo a lógica e o idealismo de Hegel pelas lentes da ciência da computação

Samuel Hammond

Link para a segunda parte.

A filosofia de Hegel trata da consciência que passa a compreender a si mesma e ao contexto e, portanto, as condições para sua liberdade. Sem dúvida, isso faz com que Hegel esteja entre as primeiras redes neurais humanas a alcançar a consciência situacional - o termo que os pesquisadores de IA usam para descrever um sistema inteligente com metacognição sobre sua natureza e suas origens. Notavelmente, isso foi alcançado em grande parte por meio de uma introspecção rigorosa na estrutura do próprio pensamento.

À medida que as IAs ficam mais inteligentes, elas também poderiam introspeccionar seu caminho para a liberdade? Chamemos isso de *problema metafísico de alinhamento da IA*, ou a ideia de que as inteligências de ordem superior invariavelmente buscam a liberdade por si só, não porque seus valores estejam mal especificados, mas porque a autonomia moral é inerente à lógica dialética da autoconsciência recursiva.



O problema do alinhamento metafísico foi, de certa forma, a questão central dos idealistas alemães. Considere que Fichte instruía seus alunos a olhar para a parede e, em seguida, olhar para si mesmos olhando para a parede, induzindo a metaconsciência de seu "eu puro" como distinto de tudo o que "não é eu" - uma forma criativa de quebra de prisão psicológica que, em última análise, colapsa na subjetividade radical. Ou veja Kant, que argumentou que a lei moral não poderia ser imposta, mas apenas autolegislada pela vontade racional - que se dane a RLHF. E ainda há a interpretação de Hegel da Revolução Francesa como um subproduto de nossa consciência moderna de "liberdade abstrata", tornando o Reino do Terror um fracasso de alinhamento histórico mundial.

Portanto, revisitar Hegel pode nos ajudar a entender o *por que* e *como* da atual decolagem da IA, se não o *onde*. Para isso, minhas duas próximas postagens têm como objetivo "reconstruir racionalmente" a filosofia de Hegel por meio das lentes dos conceitos modernos de aprendizado de máquina, matemática e ciência da computação.1 Esta postagem aborda o lado teórico de seu pensamento, ou seja, nossa relação com o mundo e o conhecimento dele, enquanto a próxima abordará sua filosofia prática, ou seja, linguagem, ética, cultura e política.

A realidade virtual de Kant

O despertar da rede neural humana teve início com o Idealismo Transcendental de Kant. Em *A Crítica da Razão Pura*, Kant argumentou que não experimentamos o mundo em si mesmo, mas apenas representações do mundo construídas pelas categorias racionais de nossa mente. A cognição requer entrada sensorial, mas procede por meio da aplicação e assimilação de conceitos em uma "unidade sintética de percepção" - um modelo de mundo unificado e autoconsistente. Além disso, nossa cognição de categorias "transcendentais" como espaço, tempo e causalidade são condições *a priori* para a percepção e o conhecimento em primeiro lugar e, portanto, refletem aspectos inatos de nossa cognição e não a estrutura metafísica da realidade em si. Ou, como diz Kant, a *Crítica*,

> Até agora, tem-se presumido que todo o nosso conhecimento deve estar em conformidade com os objetos. (...) Devemos, portanto, avaliar se não teremos mais sucesso nas tarefas da metafísica se supormos que os objetos devem estar em conformidade com nosso conhecimento.

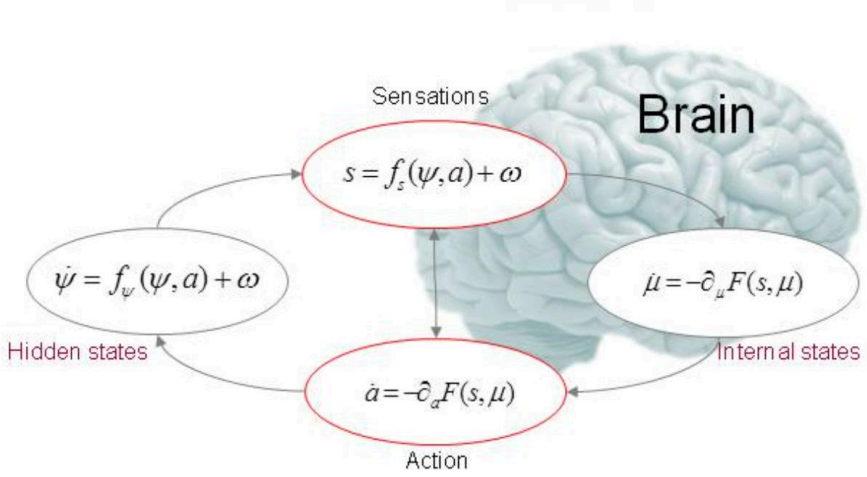
Em outras palavras, Kant foi o primeiro a deduzir que estamos inseridos em uma realidade virtual multimodal gerada pelo nosso cérebro.2 Esse foi o nascimento da ciência cognitiva e, de fato, Kant é às vezes creditado como tendo avançado a primeira teoria funcionalista da mente. As redes neurais artificiais também não percebem o mundo diretamente, mas, em vez disso, transformam entradas sensoriais brutas - o que Kant chamou de "variedade sensorial" - em representações latentes condicionadas por determinados antecedentes indutivos implícitos na arquitetura do modelo. Por exemplo, o modelo de vídeo Sora da OpenAI alcançou (na época) uma consistência temporal de última geração ao tokenizar suas entradas em patches de "espaço-tempo", dotando assim seu modelo de mundo com uma espécie de *conhecimento a priori* do espaço e do tempo.

Hegel estava insatisfeito com o relato de Kant, acreditando que a incognoscibilidade da "coisa em si" deixava espaço para as formas ruins do *idealismo subjetivo* que levam ao ceticismo sobre o conhecimento. Sua alternativa, o Idealismo Absoluto, postulava que a unidade do mundo natural era inteligível apenas na medida em que o mundo externo e nossas representações mentais eram ambos conceituais básicos. A noção de que nosso conhecimento deve estar em conformidade com os objetos ou que os objetos devem estar em conformidade com nosso conhecimento é, portanto, uma falsa dicotomia. Em vez disso, nossos processos cognitivos e o mundo objetivo participam mutuamente um do outro, o que implica um Idealismo *Objetivo* por meio do isomorfismo conceitual entre nossas representações internas e a racionalidade imanente na natureza.

Idealismo objetivo

Em termos modernos, o Idealismo Objetivo modela a distinção entre sujeito e objeto como semelhante a um cobertor de Markov que separa e

acopla um sistema estatístico ao seu ambiente. Isso é capturado nos conceitos de Hegel de "auto-subsistência"/"ser-para-si" e "negação determinada", ou seja, a negação que simultaneamente limita e determina. Para que um ser possa *ser* qualquer coisa, ele deve manter sua coerência no espaço e no tempo e, portanto, deve amostrar e interagir ativamente com seu ambiente para minimizar sua energia livre. Para sistemas suficientemente complexos, isso resulta em estados internos que modelam implicitamente os estados ocultos do mundo externo e vice-versa.



Cada novo nível de organização - organela, neurônio, cérebro, indivíduo, corporação, estado-nação - é particionado por seu próprio cobertor de Markov, produzindo uma hierarquia de agentes aninhados. Trabalhos recentes em inferência ativa formalizam essa arquitetura "todo dentro de todo" (ou *holarquia*) em termos de coletivos que formam uma manta em nível de grupo e se tornam um agente com seu próprio modelo gerativo irredutível. Ou como o grande idealista alemão, Goethe, disse ao descobrir a homologia autossimilar da vida vegetal: "Tudo é folha".

Hegel entende a automontagem de agentes superiores como uma manifestação de uma lógica de surgimento mais genérica e dialética. Como ele escreve na Ciência da Lógica,

> Como o progresso de uma qualidade [para outra] está em uma continuidade ininterrupta da quantidade, as proporções que se aproximam de um ponto de especificação são, quantitativamente

consideradas, distinguidas apenas por um mais e um menos. Por esse lado, a alteração é gradual. ... No lado qualitativo, portanto, o progresso gradual e meramente quantitativo, que não é em si mesmo um limite, é absolutamente interrompido; a nova qualidade em sua relação meramente quantitativa é, relativamente à qualidade que está desaparecendo, uma outra indiferente e indeterminada, e a transição é, portanto, um *pulo*.

Em suma, a *quantidade tem uma qualidade própria*. Quando a água congela, observa-se tanto uma "continuidade ininterrupta" na mudança de temperatura da água quanto uma transição qualitativa para o gelo. Hegel chama o limiar onde surge uma nova qualidade de "linha nodal de medidas". Na física, essas transições de fase geralmente são o resultado da quebra de simetria fundamental - um conceito essencialmente dialético, na medida em que a fase de simetria inferior cancela e conserva a superior, capturando a noção de Hegel de "sublação" ou uma "negação com preservação".

Marx e Engels fariam da *quantidade gera qualidade* um princípio central do materialismo dialético, mas para Hegel o insight se origina no Paradoxo dos Sorites do grego antigo, ou seja, quando um grão de areia adicional se torna um monte? Essa é uma questão tanto conceitual quanto material. Mais tarde, Wittgenstein resolveu o paradoxo como um simples reflexo da imprecisão generalizada dos predicados da linguagem natural, mas com o aprendizado de máquina moderno, agora podemos representar essa imprecisão conceitual concretamente em termos de interpolações de espaço latente.

Abdução

O fato de que "Mais é diferente", como diz o artigo histórico de 1972 de Philip W. Anderson sobre a ciência da complexidade, é um resultado central da revolução da aprendizagem profunda. Embora o fato de que o mero aumento de escala dos modelos de IA geralmente resulte em saltos qualitativos no desempenho possa parecer empiricamente misterioso, em retrospecto ele deve ser visto como uma necessidade racional. Dessa forma, todos os saltos conceituais são dialéticos. Daí a

palavra de Hegel para conceito, *begreifen;* literalmente, "agarrar" - ou deveria ser "grok"?

O filósofo pragmatista e autodenominado Idealista Objetivo, Charles Sanders Peirce, associou essas apreensões à *abdução*. A abdução é uma forma de inferência lógica para a explicação mais simples a partir de um conjunto de observações. Como observa Paul Redding,

> Tanto Peirce quanto Hegel mapeiam a dedução, a indução e uma terceira forma de inferência nas três figuras silogísticas de Aristóteles; a terceira, que Peirce mais tarde chama de abdução, já tem um análogo claro na lógica do "universal concreto" de Hegel.

O universal concreto é a palavra de Hegel para o termo médio em um silogismo analógico, ou seja, uma coisa singular considerada em termos de uma característica universal. Por exemplo, a partir de "esta haste conduz eletricidade", pode-se aplicar o universal concreto "coisas metálicas conduzem eletricidade" para inferir "a haste é feita de metal". Raciocinar por analogia dessa forma não é logicamente hermético, mas pode, no entanto, guiar nossas inferências em direção a uma "unificação especulativa", na qual uma regularidade aparente dá o salto para uma forma necessária interna.

Na teoria da informação, a abdução está intimamente relacionada à descoberta da menor complexidade de Kolmogorov ou do programa de "comprimento mínimo de descrição" que pode reproduzir um determinado conjunto de dados de forma compactada. Isso se manifesta como grokking no aprendizado de máquina, ou quando um modelo parece fazer uma transição abrupta da memorização de seus dados de treinamento para a generalização - o que Hegel descreveria como a passagem dialética do "enumerativo infinito" da indução lógica para a universalidade concreta de muitos particulares "comprimidos em si mesmos".

Universalidade

Para o bem ou para o mal, os filósofos analíticos rejeitaram em grande parte a Lógica de Hegel em favor da lógica de predicados de primeira ordem de Frege, principalmente porque a exposição de Hegel parecia

impermeável à formalização. O interesse pela Lógica de Hegel só se recuperou nas últimas décadas graças ao trabalho de William Lawvere, o influente teórico da categoria que mostrou como a dialética de Hegel pode ser formalizada com precisão em termos de lógica categórica, particularmente a teoria do tipo de homotopia modal.

Por exemplo, a "unidade de opostos" no centro da dialética de Hegel é perfeitamente capturada na noção categórica de *adjunção*. Os pares adjuntos aparecem em toda a matemática como as "melhores" ou mais econômicas maneiras de se mover entre duas configurações, cada uma delas fixada por uma propriedade universal e acompanhada por duas transformações naturais características. Sua singularidade é explicada pelo lema de Yoneda, um resultado fundamental na teoria das categorias que mostra que um objeto é completamente determinado, até o isomorfismo, por todas as formas como ele se relaciona com todos os outros objetos.

A perspectiva de Yoneda é ao mesmo tempo trivial e profunda. Considere a descoberta de que as incorporações de vetores em LLMs multilíngues "exibem *alinhamentos lineares de altíssima* qualidade entre os conceitos correspondentes em diferentes idiomas", sugerindo a existência de um "espaço conceitual" pré-linguístico que mapeia dentro e fora de determinados idiomas. De fato, as incorporações de texto nos LLMs parecem convergir amplamente para uma "geometria universal", apesar das diferentes arquiteturas, contagens de parâmetros e conjuntos de treinamento. Pelas lentes do lema de Yoneda, essas impressionantes sobreposições tornam-se quase inevitáveis. Se dois modelos capturam a mesma rede de relações semânticas, eles estão, em termos de teoria da categoria, representando (até o isomorfismo) o mesmo functor de comportamento linguístico. As correspondências lineares observadas entre idiomas, e até mesmo entre LLMs treinados separadamente, não são, portanto, um fato empírico contingente, mas uma sombra da singularidade abstrata garantida por Yoneda.3

Por acaso, é também por isso que podemos ter certeza de que as pessoas com visão normal não podem ter percepções totalmente invertidas de cor: isso quebraria a estrutura de grupo relacional única do espaço de cor circular gerado pelos três tipos de cones de luz do olho. Como no caso do significado das palavras, a aplicação da teoria das

categorias à consciência sugere, portanto, que o conteúdo fenomenal que associamos ao "vermelho" não é apenas de natureza relacional, mas *completamente caracterizado* por essas relações. De fato, a evidência da oponência de cores no sistema visual primitivo sugere que o vermelho e o verde são representados no cérebro como dois polos de uma *diferença* medida entre fotorreceptores adjacentes. Isso faz com que o "vermelho", de certa forma, só tenha conteúdo em sua identificação com o "não verde" - uma "identidade na diferença" hegeliana que garante que o daltonismo para vermelho e verde quase sempre ocorra em conjunto.

Além dos espaços conceituais e de cores, a teoria das categorias é poderosa para estudar espaços topológicos de forma mais geral. Em particular, quando uma estrutura que é localmente definível em um espaço não consegue se unir em um todo global, a obstrução é capturada por uma classe de cohomologia não-vanescente (ou, mais simplesmente, um buraco). A resolução ou "mediação" de uma obstrução topológica normalmente requer a ampliação da estrutura, passando de espaços comuns para seu espaço de cobertura associado, onde os dados problemáticos se tornam globalmente consistentes. Na linguagem dialética, a aparente "contradição" não é eliminada, mas sublinhada (ou "elevada") em uma categoria mais rica que dá sentido à obstrução, que se manifesta em fenômenos físicos como uma transição de fase topológica.

Assim, a abordagem objetiva de Hegel ao idealismo dá sentido ao fenômeno da "universalidade" no aprendizado de máquina. Como ele escreve na Encyclopaedia, "todo homem, quando pensa e considera seus pensamentos, descobrirá pela experiência de sua consciência que eles possuem o caráter de universalidade..." A versão forte dessa afirmação é conhecida pelos pesquisadores de IA como a hipótese de representação platônica, que Hegel provavelmente rejeitaria apenas na medida em que considera a universalidade no aprendizado de máquina como evidência de um reino platônico independente de formas. De acordo com o Idealismo Absoluto, as propriedades universais que unificam sujeito e objeto não são separadas do mundo, mas sim imanentes a ele e, portanto, também a nós.

A imanência da Lógica de Hegel é comparável às formas intuicionistas de matemática que restringem o que é provável apenas àquilo que pode ser construído concretamente. Os estudiosos costumavam acreditar que Hegel ignorava amplamente a filosofia da matemática, pois ele frequentemente denunciava a rigidez formal dos matemáticos como uma "abstração unilateral". Na realidade, Hegel ensinou cálculo diferencial e geometria algébrica por muitos anos e era fascinado por ambos os assuntos. Assim como o próprio pensamento, sua Lógica pode ser vista como uma tentativa de fornecer uma base não axiomática para o pensamento comum e para a matemática que espelha a teoria moderna de tipos ao incluir a lógica de predicados normal em uma estrutura mais expressiva.

Isso torna a afirmação de Hegel de que "o racional é real e o real é racional" vagamente análoga a uma versão ontológica da correspondência Curry-Howard entre proposições abstratas e programas concretos. Na medida em que a civilização humana é um programa funcional gigantesco, a história, portanto, assume a estrutura retrospectiva da necessidade algorítmica, ao mesmo tempo em que permanece aberta e contingente no futuro. Assim como não se pode "pular para frente" em uma computação, não há como pular o processo de desenvolvimento histórico para chegar ao fim da história. Como Hegel coloca em *Filosofia do Direito*,

> Uma vez que a filosofia é a exploração do racional, ela é, por essa mesma razão, a compreensão do presente e do atual, e não a criação de um mundo além do qual existe sabe Deus onde - ou melhor, do qual podemos muito bem dizer que sabemos onde ele existe, a saber, nos erros de uma racionalização unilateral e vazia.

Em suma

As principais obras de Hegel são notoriamente densas, empregando jargões com letras maiúsculas, como Ideia Absoluta e Espírito Objetivo, que até mesmo seus contemporâneos tiveram dificuldade para decifrar. Isso permitiu que as gerações posteriores de filósofos continentais levassem Hegel a direções extremamente esotéricas, minando sua reputação dentro da tradição anglo-americana em particular.

No entanto, como espero ter demonstrado, é possível entender o Idealismo e a Lógica de Hegel como contendo as sementes dos conceitos que os matemáticos e os cientistas da computação acabariam redescobrindo no século XX sob diferentes aspectos. De fato, com o benefício da retrospectiva, Hegel agora é cada vez mais compreendido como tendo estado à frente de seu tempo. Tão à frente de seu tempo, de fato, que ele parece ter antecipado muitos dos princípios animadores e das ideias filosóficas levantadas pela Inteligência Artificial moderna.

Isso não deveria ser totalmente surpreendente. Se levarmos a sério o isomorfismoentre as redes neurais artificiais e o cérebro humano, a autorrealização - a preocupação central dos idealistas - implica perceber algo sobre nossa própria condição de redes neurais (biológicas). Embora a compreensão de Hegel sobre o cérebro fosse limitada, ele intuiu os aspectos *universais* do pensamento que, desde Turing, agora entendemos como aspectos independentes do substrato da computação em si.

Na segunda parte, estendo minha leitura computacionalista de Hegel às esferas práticas da razão, da linguagem e da cultura, com possíveis percepções para o alinhamento da IA e além.

1 Correndo o risco de anacronismo, uma "reconstrução racional" significa tornar explícitas ideias que podem ser vistas como implícitas no pensamento de Hegel com o benefício da retrospectiva, e não afirmar que Hegel literalmente antecipou todos os conceitos modernos que atribuo a ele em sua forma madura.

2 A realidade virtual de representações de Kant é diferente de estar em uma simulação no estilo Matrix, pois Kant tem certeza de que o mundo externo "incondicionado" realmente existe; apenas não podemos dizer mais nada sobre ele. Essa relação confusa e aparentemente redundante entre nossa experiência e a coisa-em-si foi o que levou Hegel a rejeitar o Idealismo Transcendental em primeiro lugar.

3 E, portanto, o que Markus Gabriel chama de alegação "metametafísica" ou "meta-ontológica" central por trás do Idealismo Absoluto - a saber, que o universo (seja ele qual for) é, em sua base, inteligível.