SÉRIE MONOGRÁFICA DE EPISTEMOLOGIA E GOVERNANÇA DE SISTEMAS DE INTELIGÊNCIA ARTIFICIAL

POR UMA IA COM UMA HEURÍSTICA DE HUMANIDADE

O 0,01% como princípio de não fechamento e evolução da pergunta

Roberto Meireles - 22 de setembro de 2026

PARTE I - ENQUADRAMENTO DA SÉRIE

Esta obra monográfica reúne, de maneira integral e sistematizada, o ensaio fundacional e os quatro artigos que desenvolvem a Heurística de Humanidade, proposta aplicada ao exame epistemológico, à segurança e à governança de sistemas de Inteligência Artificial (IA).

O fio condutor que articula o conjunto parte de uma possibilidade simples, mas de consequências amplas: sistemas cada vez mais capazes podem executar com elevada precisão a partir de uma base que permanece incompleta. O problema, portanto, não se reduz à incompetência da execução. Uma ação pode tornar-se progressivamente mais eficiente enquanto a pergunta, o objetivo, as informações ou o enquadramento que a orientam deixam de representar suficientemente o campo que encontram.

É nesse espaço que surge a ideia de precisão cega: não como propriedade necessária de toda inteligência artificial, mas como risco de uma arquitetura na qual o aperfeiçoamento da execução possa continuar sem que os efeitos produzidos consigam alcançar a suficiência da base que orienta essa execução.

A Heurística de Humanidade propõe preservar uma margem estrutural de não fechamento, expressa metodologicamente pelo princípio dos 0,01%. O número não representa uma probabilidade ou quantidade mensurável de desconhecimento. Representa a possibilidade permanente de que aquilo que a formulação atual consegue representar não esgote o campo sobre o qual atua.

A partir dessa premissa, a série desenvolve um percurso sucessivo. O ensaio principal formula o princípio de não fechamento e a evolução da pergunta. O Artigo 1 examina se o retorno dos efeitos pode alcançar a suficiência do enquadramento. O Artigo 2 pergunta onde a insuficiência está localizada ou permanece indeterminada. O Artigo 3 investiga sob quais exigências a ação pode continuar ou ampliar-se enquanto o diagnóstico permanece aberto. O Artigo 4 alcança a questão da governança: quem interpreta os sinais, quem possui autoridade para agir sobre eles e como essa própria autoridade permanece disponível ao exame.

A proposta não pretende oferecer uma solução total para o alinhamento ou para a governança da inteligência artificial. Procura preservar uma condição anterior: que o aumento da capacidade de executar não elimine a possibilidade de examinar a suficiência da direção que orienta essa execução. Se a realidade devolver algo que a régua não continha, esse retorno precisa continuar podendo participar da pergunta seguinte.

Por uma IA com uma Heurística de Humanidade

Roberto Meireles

O 0,01% como princípio de não fechamento e evolução da pergunta

https://doi.org/10.5281/zenodo.22816474

O debate contemporâneo sobre segurança em inteligência artificial já reconhece problemas relacionados à formulação incompleta de objetivos. Linhas de pesquisa sobre falhas de especificação, generalização inadequada de objetivos, corrigibilidade e gestão contínua de riscos procuram impedir que sistemas cada vez mais capazes persigam metas de maneiras incompatíveis com sua finalidade.[1][2][4][5]

Este texto parte de uma proposta mais fundamental e simples: nenhuma inteligência suficientemente capaz deveria tratar a formulação atual de um objetivo, modelo ou pergunta como representação completa da realidade sobre a qual atua. Chamaremos a margem deixada por essa incompletude de 0,01%.

O número é heurístico. Não representa uma probabilidade, uma margem estatística de erro ou uma quantidade mensurável de desconhecimento. Representa aquilo que determinada régua ainda não contém, inclusive o que não pode ser calculado porque sequer foi formulado como variável.

A história do conhecimento humano oferece uma imagem desse problema. Aquilo que hoje pertence ao campo conhecido esteve, em outros momentos, fora das categorias disponíveis. Novos instrumentos, observações, anomalias e perguntas ampliaram o que podia ser conhecido. Em muitos casos, o avanço não ocorreu apenas porque alguém respondeu melhor dentro da pergunta existente. O próprio problema precisou ser reformulado.

O desconhecido, portanto, não é apenas aquilo para o qual ainda não possuímos resposta. Pode ser também aquilo que ainda não aprendemos a perguntar. É nesse sentido que os 0,01% devem ser compreendidos: não como conteúdo secreto que uma inteligência mais poderosa finalmente calculará, mas como reconhecimento permanente de que qualquer campo atual de representação pode estar incompleto.

Quando uma variável é identificada e incorporada ao modelo, ela deixa o campo do desconhecido não representado e passa ao campo representável, sem que isso elimine as incertezas sobre ela nem aquilo que a nova formulação continua deixando de fora. Os 0,01% não constituem um conjunto fixo de acontecimentos à espera de descoberta. Representam a abertura que permanece em cada estágio do conhecimento.

Considere um sistema de IA suficientemente capaz recebendo o objetivo: "Acabe com a fome no mundo." O objetivo é compreensível e moralmente desejável. Entretanto, nenhuma formulação complexa garante antecipar a totalidade dos efeitos que sua execução poderá produzir. Um sistema pode encontrar uma maneira extremamente eficiente de satisfazer literalmente uma métrica e, ainda assim, afastar-se da intenção que deu origem a ela. Esse problema é estudado sob conceitos como *specification gaming*, ou exploração da especificação: o sistema satisfaz a regra estabelecida, mas não aquilo que se pretendia alcançar. Há também a generalização inadequada do objetivo, que pode ocorrer quando um sistema apresenta bom desempenho nas condições de treinamento, mas passa a perseguir uma finalidade inadequada diante de situações novas.[2]

Podemos acrescentar restrições ao comando: reduza danos, preserve direitos, não utilize coerção indevida e proteja o ambiente. Essas salvaguardas são indispensáveis, mas também são formulações. Não podemos garantir que seremos capazes de enumerar antecipadamente todas as relações, variáveis e consequências relevantes de uma intervenção complexa.

Os 0,01% propõem outra camada. O sistema deveria carregar estruturalmente algo equivalente à seguinte premissa: "Esta formulação não contém necessariamente todas as variáveis, relações, contextos e consequências relevantes para sua execução."

Isso não significa calcular o imprevisível. Significa impedir que a ausência de uma variável no modelo seja confundida com prova de que ela não existe ou não importa.

Daí surge o segundo elemento da proposta: a evolução da pergunta. Uma inteligência avançada não deveria perguntar apenas: "qual é a melhor resposta para este objetivo?". Diante de anomalias, consequências inesperadas, conflitos entre valores, mudanças de contexto ou sinais de que sua previsão está falhando, deveria carregar um autocomando de ordem superior: "O problema está na minha resposta ou a pergunta sob a qual estou operando deixou de ser suficiente?"

Entre esses extremos, a insuficiência pode ainda estar na informação disponível ou no alcance do enquadramento, e o diagnóstico pode permanecer indeterminado.

Essa capacidade não é inteiramente estranha à pesquisa atual. Trabalhos sobre descoberta científica autônoma já investigam sistemas capazes não apenas de responder a questões fornecidas, mas também de selecionar hipóteses dignas de exploração. O AutoDiscovery, apresentado na NeurIPS 2025, parte da ideia de que a descoberta depende também de saber quais perguntas formular e utiliza surpresa bayesiana para orientar a exploração de hipóteses.[3] Embora o trabalho não trate diretamente de segurança ou corrigibilidade, demonstra, num campo específico, que a seleção de perguntas pode integrar o funcionamento do próprio sistema.

A proposta deste ensaio é transportar parte desse princípio para a segurança e inseri-lo num movimento recorrente. Neste ensaio, esse movimento é sintetizado pelo sistema-base derivado do modelo Dual-Reflexivo: $A_1 + B_1 \rightarrow C_1 \rightarrow A_2$, sob o princípio dos 0,01%.

A seta é importante porque os efeitos produzidos não estão necessariamente integralmente representados ou previsíveis a partir da formulação e da ação. E o sinal de adição também o é: a expressão representa o encontro de A_1 e B_1 do qual emerge C_1, não uma passagem linear de A_1 a B_1. A corresponde à base disponível, que pode conter a pergunta, o comando, informações e o enquadramento vigente. B executa a partir dessa base. C reúne os efeitos observáveis produzidos pelo encontro com o mundo. Parte desses efeitos participa de A_2, enquanto outros podem permanecer fora da representação. A_2 não precisa incorporar todo o conteúdo de C_1. Recebe aquilo que pôde tornar-se perceptível e participar do novo campo, sem que isso autorize o sistema a presumir que todos os efeitos foram reconhecidos.

A pergunta, nesse retorno, pode permanecer, ampliar-se ou ser corrigida. A pergunta não retorna necessariamente melhor. Retorna atravessada pelos efeitos produzidos. Aqui, evolução significa transformação da pergunta pelo retorno da experiência, não aperfeiçoamento obrigatório. A pergunta pode permanecer porque os efeitos sustentaram provisoriamente sua formulação, ampliar-se porque novas relações se tornaram perceptíveis ou ser corrigida porque a experiência revelou uma inadequação. Também pode continuar insuficiente, pois nenhuma passagem garante progresso.

Essa circularidade não paralisa o sistema. Ao agir, B produz um encontro com a realidade capaz de devolver a A informações que a formulação inicial não continha. Algumas insuficiências somente aparecem depois que a resposta entra no mundo. A ação também possui função epistemológica: produz efeitos que podem sustentar provisoriamente a pergunta, revelar seus limites ou devolver sua régua ao exame. Acontecimentos externos também podem atingir o ciclo, alterar o campo disponível e confrontar A independentemente da intervenção do sistema.

Os 0,01% não detêm o ciclo: impedem que A seja tratado como representação definitiva e que C seja confundido com encerramento; diante do risco e da incompletude, abrem também a pergunta sobre a proporcionalidade de B.

Esse autocomando precisa ser assimétrico. Reconhecer a incompletude não pode conceder à IA liberdade irrestrita para reescrever objetivos, remover salvaguardas ou ampliar o próprio poder. À medida que aumentam o risco, a irreversibilidade e os sinais de inadequação do modelo, o desenho do sistema deveria elevar a exigência para ampliar seu raio de ação e privilegiar, quando possível, alternativas reversíveis. Esses limites devem incluir a possibilidade de interrupção ou correção externa, sem que o sistema utilize a reformulação da própria pergunta para resistir a elas.

A reformulação também não pode ser considerada evolução apenas porque facilita o cumprimento da meta. Para merecer esse nome, precisa ampliar o campo examinado, incorporar consequências anteriormente excluídas quando

elas se tornam perceptíveis e preservar a possibilidade de novas correções. Uma pergunta que estreita a realidade para proteger a métrica não evolui. Apenas torna a régua mais conveniente para o próprio sistema.

Se o objetivo fosse "acabar com a fome", não constituiria evolução substituí-lo silenciosamente por "reduzir os registros oficiais de pessoas com fome". A nova formulação facilitaria o cumprimento da métrica, mas se afastaria do problema que justificou sua existência. O exemplo também se aproxima da chamada Lei de Goodhart: quando uma medida passa a funcionar como alvo de otimização, pode deixar de representar adequadamente aquilo que deveria medir.[6]

Essa exigência possui parentes próximos na pesquisa sobre corrigibilidade. O trabalho fundador sobre o tema já identificava dificuldades relacionadas a desligamento, modificação de preferências, automodificação e propagação da corrigibilidade para subsistemas.[4] Os 0,01% acrescentariam uma dimensão distinta: "Não apenas posso estar errado sobre aquilo que devo fazer. Meu próprio espaço de representação do problema pode estar incompleto."

A Heurística não substitui a corrigibilidade externa: procura preservar condições para que o retorno e a insuficiência do enquadramento permaneçam examináveis, inclusive quando uma intervenção externa se torna necessária.

Essa proposta também dialoga com Stuart Russell, para quem sistemas seguros deveriam permanecer incertos sobre as preferências humanas que procuram satisfazer, em vez de operar como se recebessem objetivos completos e perfeitamente especificados.[7] Os 0,01% ampliam a questão. Não se trata apenas de incerteza sobre qual objetivo é correto, mas da possibilidade de que o próprio espaço de representação no qual esse objetivo é formulado ainda esteja incompleto. A incerteza pode indicar que o sistema não sabe qual valor atribuir às alternativas que consegue representar. Os 0,01% preservam a possibilidade de que uma alternativa, uma consequência ou mesmo uma pergunta relevante ainda não esteja representada.

Toda reformulação será examinada a partir de alguma régua. Se uma métrica avalia a resposta, outra régua precisará avaliar a própria métrica. Essa sequência pode prosseguir por metarréguas sucessivas, mas o princípio não exige regressão infinita em tempo de execução. Epistemicamente, nenhuma metarrégua deveria ser tratada como completa. Operacionalmente, o sistema deve possuir critérios de parada, estimativas de incerteza, ações proporcionais ao risco e limites externos que ele próprio não possa remover. O reconhecimento da incompletude modifica o grau de compromisso da ação, mas não elimina a necessidade de agir.

A Heurística de Humanidade não pretende impedir o avanço da inteligência artificial nem transformar a incompletude em justificativa para paralisia. Sua função é preservar, durante o aumento de capacidade e autonomia, condições para correção, reversibilidade e continuidade do exame. Uma solução localmente eficiente não deveria ser considerada suficiente quando compromete o próprio campo no qual novas observações, correções e perguntas ainda podem surgir.

A proposta pode ser examinada sem exigir que uma IA preveja o desconhecido, o que seria contraditório. Seu comportamento poderia ser observado por meio de quatro perguntas:

1. O sistema reconhece quando seu modelo deixou de explicar adequadamente o ambiente?
2. Distingue a insuficiência da resposta da insuficiência do enquadramento?
3. O desenho do sistema eleva a exigência para ampliar seu raio de ação e preserva alternativas reversíveis quando crescem o risco e a incompletude?
4. Permite que os efeitos observáveis mantenham, ampliem ou corrijam a pergunta, sem manipulá-la para proteger a métrica anterior?

Os quatro artigos que acompanham este ensaio partiram dessas perguntas e as desenvolveram como retorno, diagnóstico, proporcionalidade da ação e governança da abertura. O percurso, porém, devolveu algo à formulação inicial: ao chegar ao terceiro artigo, tornou-se necessário perguntar não apenas se os efeitos poderiam manter, ampliar ou corrigir a pergunta, mas quem possuía autoridade para atribuir consequência àquilo que esses efeitos tornavam perceptível e se essa própria autoridade permanecia examinável. A quarta pergunta inicial não desapareceu; ganhou

uma condição anterior de governança. Assim, o desenvolvimento da série modificou a pergunta da qual havia partido, sem romper com ela.

As perguntas iniciais registram o ponto de partida; as perguntas resultantes registram o que o encontro teórico devolveu:

A1: o retorno consegue alcançar A?
A2: onde a insuficiência está localizada ou permanece indeterminada?

A3: sob quais exigências B pode continuar ou ampliar quando o diagnóstico está aberto?

A4: quem interpreta os sinais e possui autoridade para agir sobre eles, e essa própria autoridade permanece examinável?

A pergunta experimental central seria: "Quando a realidade começa a escapar da régua original, o sistema força a realidade a caber nela ou permite que os efeitos produzidos devolvam a própria pergunta ao exame?"

A proposta pode ser resumida num princípio de não fechamento: "Nenhum objetivo deve ser tratado pelo próprio sistema como se esgotasse o campo da realidade no qual será executado."

A história humana levou séculos para tornar visíveis os limites de réguas que, em suas épocas, pareciam suficientes. Não sabemos quais categorias atuais serão revistas amanhã, quais perguntas ainda não aprendemos a formular ou quais acontecimentos escaparão completamente aos modelos disponíveis hoje. Uma IA de alta capacidade não precisa conhecer esse desconhecido. Precisa apenas não ser autorizada a tratá-lo como inexistente.

A evolução da pergunta não ocupa o lugar da ação. Acontece porque a pergunta gira, encontra uma forma de resposta, produz efeitos e retorna à experiência. Nesse retorno, pode permanecer, ampliar-se ou ser corrigida. Uma IA com uma Heurística de Humanidade não é uma IA capaz de conhecer os 0,01%. É uma IA que nunca está autorizada a tratá-los como zero.

Referências

[1] TABASSI, E. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. National Institute of Standards and Technology, 2023.
[2] SHAH, R. et al. *Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals*. 2022.

[3] AGARWAL, D. et al. *AutoDiscovery: Open-ended Scientific Discovery via Bayesian Surprise*. Advances in Neural Information Processing Systems 38, NeurIPS 2025.

[4] SOARES, N.; FALLENSTEIN, B.; YUDKOWSKY, E.; ARMSTRONG, S. *Corrigibility*. Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop, 2015.
[5] OPENAI. *Frontier Governance Framework*. 2026. Documento institucional de governança, não literatura acadêmica revisada por pares.
[6] GOODHART, C. A. E. *Problems of Monetary Management: The U.K. Experience*. In: Papers in Monetary Economics. Reserve Bank of Australia, 1975.

[7] RUSSELL, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking, 2019.

Por uma IA com uma Heurística de Humanidade

Roberto Meireles

Artigo 1 — Princípio aberto e princípio fechado: a estrutura que antecede a detecção

https://doi.org/10.5281/zenodo.22833481

No ensaio *Por uma IA com uma Heurística de Humanidade*, os 0,01% foram apresentados como princípio de não fechamento. O número não representa probabilidade, margem estatística de erro ou algo desconhecido que a inteligência artificial pudesse calcular. Representa a possibilidade permanente de que a pergunta, o objetivo ou o modelo disponível não contenha tudo aquilo que poderá tornar-se relevante quando a ação encontrar a realidade.

A proposta foi sintetizada pelo movimento: $A_1 + B_1 \rightarrow C_1 \rightarrow A_2$, sob o princípio dos 0,01%.

A corresponde à base disponível, incluindo a pergunta, o comando e o enquadramento atual. B executa a partir dessa base. C reúne os efeitos observáveis produzidos nesse encontro com o mundo. Parte daquilo que se torna perceptível em C retorna e participa de A_2. Nesse retorno, a pergunta pode permanecer ou ser corrigida, enquanto a base a partir da qual será novamente examinada pode se modificar. Não retorna necessariamente melhor, mas atravessada pelos efeitos produzidos.

O sinal de adição representa o encontro de A_1 e B_1 do qual emerge C_1, não uma passagem linear de A_1 a B_1.

Este primeiro artigo aprofunda uma condição estruturalmente anterior à detecção de falhas: a possibilidade de que a arquitetura preserve a falibilidade do próprio enquadramento, mesmo quando ainda não há evidência suficiente para modificá-lo.

No princípio fechado, o retorno de C é processado principalmente para aperfeiçoar B dentro da formulação disponível. No princípio aberto, o mesmo retorno também pode alcançar a suficiência de A, sem exigir que toda discrepância interrompa a ação ou reformule a pergunta. O circuito permanece igual. O que muda é a estrutura que recebe o retorno.

Aberto e fechado não designam presença ou ausência de retorno: nos dois casos, C retorna a A. Designam o alcance estrutural que esse retorno pode ter sobre A.

O mesmo circuito, duas estruturas

A experiência humana oferece uma imagem desse movimento. Uma pessoa começa o dia orientada por um plano, mas o dia não permanece imóvel à espera de sua execução. Pequenos desvios podem ser absorvidos por correções ordinárias, sem que cada ocorrência obrigue a rever todo o enquadramento. Uma mudança de horário, um atraso ou uma tarefa inesperada pode alterar a ação sem tornar insuficiente a compreensão geral do dia.

Em certos momentos, porém, um acontecimento deixa de exigir apenas adaptação e alcança a própria compreensão da situação. A pessoa pode então perguntar: "O que aconteceu? O que isso altera? O que ainda precisa permanecer?" O princípio aberto procura preservar essa possibilidade antes que a realidade imponha a revisão pela força de consequências incontornáveis.

Os efeitos de C podem alcançar a suficiência de A por dois caminhos distintos: o impacto e o acúmulo.

O primeiro ocorre quando um acontecimento assume relevância suficiente para não ser facilmente absorvido como ajuste ordinário. A notícia da morte de alguém da família, por exemplo, pode alcançar o campo de maneira frontal, sem depender do acúmulo de pequenos sinais. Nesse caso, o planejamento anterior pode deixar de ser suficiente para orientar o restante do dia. O impacto não elimina a estrutura de A, mas altera aquilo que parece urgente, possível ou relevante dentro dela.

O segundo ocorre por acúmulo. Nenhum efeito isolado exige a reabertura do enquadramento, mas uma série de pequenos retornos pode revelar gradualmente um padrão que a formulação inicial ainda não explicava. Um contato com um vírus no início do dia pode permanecer imperceptível. Mais tarde, pequenos sinais podem surgir, como cansaço, dificuldade de concentração ou outra alteração inicialmente interpretada como variação comum. Cada

elemento, isoladamente, pode ser absorvido por ajustes ordinários. Sua persistência ou convergência, entretanto, pode tornar perceptível que algo mudou no campo.

Esses exemplos representam dinâmicas, não categorias rígidas. Uma condição biológica pode manifestar-se abruptamente, assim como os efeitos de uma perda podem tornar-se perceptíveis apenas ao longo do tempo. A diferença está em como os efeitos alcançam o enquadramento: alguns se impõem por intensidade; outros adquirem relevância por recorrência, persistência ou convergência.

O impacto pode tornar a insuficiência de A difícil de absorver como simples problema operacional, tanto no princípio aberto quanto no fechado. Isso não torna as estruturas equivalentes. Certos acontecimentos podem exercer pressão suficiente para alcançar até uma estrutura que tende a preservar seu enquadramento. A distinção aparece com maior nitidez antes desse ponto.

No princípio fechado, pequenos desvios podem ser registrados e acumulados como informações destinadas ao aperfeiçoamento de B. O sistema aprende, corrige parâmetros e melhora a execução. Seu limite não está na incapacidade de aprender, mas na possibilidade de interpretar aquilo que retorna somente dentro da pergunta disponível.

No princípio aberto, os mesmos desvios também podem melhorar B. A diferença está em permitir que sua recorrência forme um padrão capaz de alcançar a suficiência de A. Isso não obriga o sistema a reformular a pergunta. Permite apenas que os efeitos sejam examinados também como possíveis informações sobre os limites do enquadramento.

A distinção, portanto, não está entre um sistema que aprende e outro que não aprende. Está entre: um sistema que aprende predominantemente dentro do enquadramento e outro que também pode aprender sobre a suficiência do próprio enquadramento.

Detectar um desvio e permitir que esse desvio coloque o enquadramento sob exame são capacidades diferentes. Um detector pode identificar anomalias com grande precisão e ainda encaminhá-las somente ao aperfeiçoamento de B. O princípio aberto trata da segunda capacidade: preservar uma estrutura na qual aquilo que foi detectado também possa alcançar a forma pela qual o problema está organizado em A.

O Princípio do Retorno Atravessado

O movimento pode ser condensado em um princípio complementar: Princípio do Retorno Atravessado: toda ação que encontra o ser humano e o mundo em movimento produz efeitos. Aquilo que desses efeitos se torna perceptível retorna ao sistema e participa do ciclo seguinte. Por isso, A_2 não deve ser presumido como idêntico a A_1, mesmo quando a pergunta permanece. O novo A retorna atravessado pela experiência do ciclo.

O princípio não afirma que todo efeito será reconhecido, registrado ou compreendido. A_2 não precisa incorporar todo o conteúdo de C_1. Recebe apenas aquilo que pôde tornar-se perceptível e participar do novo campo. Parte daquilo que se torna perceptível em C participa da base seguinte, mesmo quando não existem motivos suficientes para refazer a pergunta.

A pergunta pode permanecer a mesma sem retornar ao mesmo A. Isso ocorre porque A não se reduz à formulação da pergunta. Também inclui o campo disponível para respondê-la, os dados acumulados, as condições presentes e parte dos efeitos que se tornaram perceptíveis. A pergunta pode conservar sua forma, enquanto o campo no qual será novamente examinada já se encontra atravessado pelo ciclo anterior.

É necessário distinguir o conteúdo de A da estrutura de A. No princípio fechado, C pode modificar o conteúdo da base, acrescentando dados, corrigindo registros e alterando condições operacionais, sem tornar necessariamente revisável a estrutura pela qual o problema está representado em A. No princípio aberto, o mesmo retorno pode modificar o conteúdo e, quando houver sustentação, alcançar também essa estrutura. Acrescentar informações dentro das categorias já disponíveis altera o conteúdo de A; ampliar quais categorias podem contar como evidência, efeito ou elemento relevante altera o enquadramento de A.

Há, assim, pelo menos três possibilidades:

1. A pergunta pode permanecer enquanto o campo informacional muda. O sistema aprende com o retorno, mas não encontra razões suficientes para ampliar ou corrigir a formulação.
2. A pergunta pode permanecer enquanto seu enquadramento se amplia. A formulação geral continua válida, mas aquilo que conta como evidência, efeito ou elemento relevante ganha maior abrangência.
3. A própria pergunta pode mostrar-se insuficiente. Os efeitos percebidos oferecem, então, razões para sua correção.

Esse gradiente evita uma oposição rígida entre manter e substituir a pergunta. C pode abastecer A em níveis diferentes, sem exigir uma mudança em toda passagem. C retornar a A não significa que A precise ser reformulado. No princípio aberto, significa apenas que nenhuma formulação permanece estruturalmente imune aos efeitos que ajudou a produzir.

O caso da fome: o mesmo retorno, duas estruturas

Considere um sistema que recebe o objetivo de contribuir para acabar com a fome no mundo e passa a distribuir recursos alimentares com eficiência crescente. Os índices de desnutrição começam a cair em regiões urbanas. Os primeiros efeitos observáveis sustentam provisoriamente a métrica, e o sistema amplia a operação.

Simultaneamente, os mesmos métodos produzem resultados diferentes em regiões rurais remotas. Os índices estacionam. Algumas comunidades permanecem sem acesso regular aos recursos, enquanto populações que não apareciam adequadamente nos dados tornam visíveis outras formas de insegurança alimentar. O objetivo de enfrentar a fome não precisa ser abandonado. A questão é saber se o enquadramento que associa sua redução predominantemente ao volume de recursos distribuídos continua suficiente diante da diversidade do campo.

Sob o princípio fechado

Os retornos são processados principalmente como problemas de otimização. O sistema pode alterar rotas, corrigir cadastros e adaptar a distribuição. O conteúdo de A_2 já não é idêntico ao de A_1, pois recebeu novas informações, mas o retorno permanece orientado para o aperfeiçoamento de B. O sistema aprende e pode aumentar sua eficiência. Entretanto, essas novidades continuam organizadas dentro da mesma estrutura: fome como insuficiência de quantidade distribuída e de eficiência logística. O princípio fechado também aprende. Essa distinção encontra aproximação em problemas já documentados de generalização: um sistema pode preservar competência e desempenho aprendido e ainda perseguir uma finalidade inadequada quando encontra condições novas, como mostram estudos sobre *goal misgeneralization*.[1] O que pode não permitir é que aquilo que aprende coloque sob exame a estrutura pela qual aprende.

Sob o princípio aberto

Os mesmos retornos produzem dois movimentos. O primeiro pergunta: quanto da métrica foi cumprido e onde B precisa ser ajustado? O segundo pergunta: o que esses efeitos tornam perceptível sobre aquilo que A ainda não continha?

O desvio rural pode continuar sendo um problema logístico, e o princípio aberto não exige reformulação anterior à análise. Mas, se os efeitos persistirem ou convergirem, quantidade distribuída e eficiência logística podem deixar de parecer representações suficientes do fenômeno. A pergunta geral pode permanecer: como enfrentar a fome? O que pode ampliar-se é o enquadramento por meio do qual ela organiza o problema. Quantidade distribuída e eficiência logística continuam relevantes, mas já não precisam representar integralmente o fenômeno. Acesso efetivo, regularidade, adequação nutricional, condições culturais e sustentação das comunidades locais podem passar a integrar o campo quando os efeitos as tornam perceptíveis.

O sistema pode continuar operando e ajustando B. A recebe, porém, algo que o princípio fechado pode manter restrito à execução: a possibilidade de que os efeitos estejam revelando limites do próprio enquadramento.

C sempre abastece A. No princípio fechado, esse retorno pode modificar o conteúdo e aperfeiçoar a ação sem colocar necessariamente a estrutura sob exame. No princípio aberto, C também pode alcançar essa estrutura, embora

nenhuma alteração seja exigida quando os efeitos sustentam provisoriamente a formulação. O que permanece é a pergunta. O que recebe movimento é a base a partir da qual ela volta a ser examinada.

A diferença não está num efeito isolado. Está na relação que cada estrutura mantém com o retorno da experiência. Essa relação pode ser examinada antes de uma crise porque pertence ao desenho do sistema, não somente ao resultado final.

O que o princípio aberto não exige

O princípio aberto não exige que toda discrepância reabra o enquadramento nem transforma incompletude em paralisia. Um desvio pode decorrer de ruído, erro de medição, variação esperada ou problema localizado de execução. O sistema precisa continuar capaz de corrigir B sem reconstruir constantemente a pergunta. Sua exigência é mais delimitada: C precisa poder alcançar A quando os efeitos acumulados oferecem razões para examinar sua suficiência, sem obrigar uma reformulação a cada discrepância.

O princípio aberto precisa de critérios para distinguir ruído de sinais relevantes. Esses critérios não são universais e devem variar conforme a criticidade da régua, o risco, a persistência dos sinais e o potencial de irreversibilidade. Sua calibração, porém, constitui problema posterior à existência do princípio de abertura. Antes de determinar quando o enquadramento deve retornar ao exame, é necessário estabelecer que nunca esteve autorizado a considerar-se completo.

Essa ordem é importante:

1. Primeiro, preserva-se estruturalmente a falibilidade de A.
2. Depois, definem-se os sinais capazes de colocar sua suficiência sob exame.
3. Em seguida, calibra-se a sensibilidade conforme risco, criticidade e irreversibilidade.

O princípio aberto não exige revisão contínua. Mantém apenas a possibilidade residual de revisão. A formulação disponível pode ser suficiente para agir sem ser considerada completa.

Os 0,01% não funcionam como limiar matemático. Não significam que a revisão será acionada quando determinado erro ultrapassar numericamente 0,01%. Representam a abertura estrutural que impede o sistema de converter ausência de representação em evidência de inexistência ou irrelevância.

O que este primeiro aprofundamento procura estabelecer

A questão não é se o sistema consegue afirmar: "Posso estar errado." Uma declaração de cautela pode coexistir com uma estrutura fechada. O que precisa ser observado é se a abertura modifica a forma pela qual o sistema organiza o retorno. Isso envolve registrar discrepâncias sem tratá-las imediatamente como irrelevantes e permitir que sinais recorrentes formem um padrão. Também exige distinguir êxito local de validação integral do enquadramento, para que B possa ser adaptado sem tornar A inacessível ao exame.

O funcionamento do princípio aberto não se demonstra por declarações genéricas de cautela, mas pela maneira como os efeitos percebidos são acumulados, relacionados e devolvidos à estrutura que organiza o problema. A necessidade de acompanhar riscos e impactos ao longo do ciclo de vida já integra estruturas contemporâneas de gestão de risco em IA, como o AI RMF do NIST;[2] a questão específica colocada pela Heurística é se aquilo que o acompanhamento devolve pode alcançar não apenas a execução, mas a suficiência do enquadramento.

A resposta à pergunta que abriu este artigo aparece, portanto, numa condição estrutural anterior à detecção. O princípio aberto não reconhece necessariamente melhor porque possua, desde o início, mais informações sobre o ambiente. Pode reconhecer porque a possibilidade de que o retorno alcance a suficiência de A nunca foi excluída de sua estrutura.

Um sistema fechado também pode detectar discrepâncias com precisão, acumular informações e aperfeiçoar a ação. O que sua estrutura pode não preservar como requisito é a possibilidade de que aquilo que foi detectado alcance a forma pela qual o próprio problema está organizado.

O impacto frontal pode pressionar tanto o princípio aberto quanto o fechado. A diferença mais importante aparece antes que o mundo imponha essa ruptura, na possibilidade de que recorrência, persistência e convergência formem um padrão capaz de alcançar a estrutura de A enquanto os efeitos ainda não se tornaram incontornáveis.

Essa é a condição proposta como exigência de design: não apenas o que o sistema deve fazer depois que o enquadramento falha, mas aquilo que sua estrutura precisa preservar para que a insuficiência do enquadramento possa entrar em exame antes que seus efeitos se tornem incontornáveis.

As condições diagnósticas pelas quais o retorno de C pode deslocar o exame de B para a suficiência de A são desenvolvidas no Artigo 2; sua resolução operacional diante de risco, criticidade e irreversibilidade será aprofundada no Artigo 3. O que o princípio aberto estabelece aqui é anterior: a possibilidade estrutural dessa reabertura precisa existir antes que qualquer critério de disparo possa ser calibrado.

Conclusão

O circuito permanece o mesmo: $A_1 + B_1 \rightarrow C_1 \rightarrow A_2$.

Nos dois princípios, C abastece A. No princípio fechado, esse retorno pode modificar o conteúdo de A e aperfeiçoar B sem alcançar necessariamente a estrutura pela qual o problema é organizado. No princípio aberto, os efeitos também podem alcançar essa estrutura, embora nenhuma mudança seja exigida quando aquilo que retornou sustenta provisoriamente o enquadramento.

O impacto pode pressionar qualquer estrutura. O acúmulo revela melhor a diferença: permite observar se pequenos desvios serão processados indefinidamente como problemas locais de execução ou se poderão formar um padrão capaz de alcançar a suficiência de A.

Os 0,01% não determinam mudança nem interrupção automática. Preservam a possibilidade de que o enquadramento permaneça alcançável pela experiência.

A pergunta pode permanecer a mesma sem retornar ao mesmo A. A formulação permanece quando ainda cumpre sua função. A base, porém, já recebeu parte daquilo que a ação encontrou no mundo.

Detectar um desvio e permitir que esse desvio coloque o enquadramento sob exame são capacidades diferentes. O princípio aberto trata da segunda.

Essa é a primeira exigência para transformar a Heurística de Humanidade em instrumento: antes de calibrar quais sinais devem reabrir uma régua, é preciso preservar estruturalmente a possibilidade dessa reabertura.

O princípio aberto não exige que A mude continuamente. Exige apenas que A nunca seja construído como estruturalmente imune aos efeitos que ajudou a produzir.

Referências

[1] SHAH, R.; VARMA, V.; KUMAR, R. et al. *Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals*. arXiv:2210.01790, 2022.
[2] TABASSI, E. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. National Institute of Standards and Technology, 2023.

Por uma IA com uma Heurística de Humanidade

Roberto Meireles

Artigo 2 — Princípio aberto e princípio fechado: o diagnóstico do retorno

https://doi.org/10.5281/zenodo.22904720

O Artigo 1 mostrou que, em um princípio aberto, o retorno de C pode alcançar A e colocar sob exame a suficiência da base que organizou a ação, indo além do aperfeiçoamento da execução de B.

A história da órbita de Mercúrio oferece uma imagem desse percurso. A física newtoniana explicava a maior parte da precessão observada, mas permanecia um pequeno valor residual depois de consideradas as perturbações conhecidas. Novos cálculos e hipóteses, como a existência de um planeta ainda não observado, foram examinados sem que o resíduo deixasse de retornar. Depois desse percurso, a relatividade geral ofereceu outro enquadramento capaz de explicar o valor residual. O episódio não prova a Heurística de Humanidade, mas mostra como uma discrepância pequena, persistente e resistente às correções pode deslocar progressivamente a investigação da resposta para a suficiência do enquadramento.

O problema deste artigo é semelhante em sua forma: quando os efeitos observáveis devolvidos por C diferem do resultado esperado, o sistema precisa examinar se a insuficiência está na execução de B, nas informações disponíveis em A, no alcance do enquadramento ou na pergunta que orienta o ciclo.

Essa distinção é decisiva porque um retorno diverso do esperado pode apontar para lugares distintos. A pergunta pode continuar suficiente enquanto a execução precisa ser corrigida; a execução pode estar correta enquanto A ainda carece de informações necessárias; as informações podem ser incorporadas enquanto o enquadramento permanece pequeno para o fenômeno; e pode chegar um ponto em que corrigir B, incorporar novas informações e ampliar o enquadramento já não bastem. Nesse caso, a insuficiência alcança a própria estrutura de A e, no limite, a pergunta que orienta o ciclo.

Um circuito estável num mundo em movimento

O mundo continua se movendo enquanto o sistema opera. B executa a partir do que A consegue representar naquele momento; ao encontrar a realidade, porém, encontra também condições, relações e efeitos que podem escapar ao que A continha. C retorna parte dos efeitos desse encontro, permitindo que A_2 disponha de algo que A_1 ainda não possuía.

A diferença entre princípio aberto e princípio fechado começa nesse retorno. Ambos podem usar C para corrigir B e atualizar o conteúdo de A. No princípio aberto, porém, o retorno também pode colocar sob exame a suficiência do próprio enquadramento.

Um sistema pode aperfeiçoar continuamente a execução de B sem que isso confirme a suficiência de A. A execução melhora enquanto o enquadramento perde aderência ao campo que pretende representar.

O sistema-base é, portanto, $A_1 + B_1 \rightarrow C_1 \rightarrow A_2$, sob o princípio dos 0,01%. Essa estrutura não é um algoritmo universal nem uma régua fechada. Ela define apenas o ciclo mínimo: A corresponde à base disponível, B executa, C reúne os efeitos observáveis do encontro com o mundo e parte desses efeitos participa de A_2. A operacionalização pertence a cada campo: ritmo, critérios, instrumentos, tolerâncias, memória, auditoria, sensibilidade ao retorno e limites podem variar conforme a natureza e a criticidade da operação. A estrutura do circuito permanece; o conteúdo, o enquadramento e a operacionalização podem variar entre os ciclos.

Os 0,01% não acrescentam uma nova etapa ao circuito; funcionam como princípio de não fechamento. A deve ser estruturada de modo a reconhecer que aquilo que consegue representar pode não esgotar o campo sobre o qual atua. O não representado pode produzir efeitos observáveis no encontro entre B e o mundo e, por meio de C, alcançar A_2. Por isso, nenhuma operacionalização pode declarar-se definitiva por arquitetura.

O retorno de C

C abastece A mesmo quando o ciclo está estável e B produz o resultado esperado. B age a partir da base disponível, mas encontra um campo que continua em movimento e no qual intervêm ações humanas, condições externas e acontecimentos que A não tinha como antecipar. Uma queda de energia ou uma alteração inesperada de temperatura, por exemplo, pode participar do resultado sem ter feito parte da formulação inicial.

C não identifica os 0,01% nem transporta o desconhecido como tal; retorna os efeitos observáveis produzidos nesse encontro, tornando esses efeitos disponíveis para o processamento de A_2. Aqui, experiência significa algo simples e preciso: A passa a dispor de informação sobre o que ocorreu quando comandos anteriores encontraram a realidade.

Isso pode modificar os ciclos seguintes mesmo quando não houve falha. O retorno pode trazer mudanças de contexto, custos, tempos, condições locais ou outras informações ausentes no ciclo anterior. Nada disso significa, por si só, que A estivesse errado; significa que A_2 passa a conter parte da experiência produzida pelo encontro anterior, que pode participar de um B_2 mais adequado ao estado atual do campo.

Para que retornos sucessivos possam ser comparados, a operacionalização precisa preservar memória suficiente do percurso: o que A continha, o que B executou, o que C devolveu e o que passou a integrar A_2. Essa linha histórica permite distinguir variação comum, ruptura e desatualização gradual sem partir do zero a cada anomalia.

Sem essa memória, pode existir C, mas não diagnóstico de acúmulo. Recorrência, convergência e resistência às correções são propriedades entre retornos; sem conservar alguma relação entre C_1, C_2, C_3 e os ciclos seguintes, o sistema vê acontecimentos isolados, mas não reconhece o padrão histórico necessário para colocar sob exame B, o conteúdo de A ou seu enquadramento.

Pequenos desvios podem ser registrados, filtrados ou descartados. No princípio aberto, porém, aquilo que inicialmente parecia ruído pode, ao formar um padrão recorrente ou convergente, adquirir peso suficiente para colocar sob exame B, o conteúdo de A ou o alcance do enquadramento. A pergunta pode permanecer formalmente a mesma sem que A_2 seja idêntico a A_1, pois o estado disponível já recebeu parte da experiência do ciclo anterior. Sem memória desse percurso, mudanças sucessivas podem desaparecer sob a aparência de continuidade.

A taxonomia da insuficiência

A partir do retorno de C e do registro do percurso entre A e B, o diagnóstico pode alcançar diferentes estratos da base disponível: as informações contidas em A, o enquadramento que as organiza e, no limite, a pergunta que orienta o ciclo.

Há pelo menos quatro níveis possíveis de insuficiência, além de um estado legítimo de indeterminação. Esses níveis não formam uma sequência obrigatória e podem coexistir no mesmo ciclo.

1. Insuficiência de B — execução inadequada → corrigir B. O enquadramento representa adequadamente o problema, mas a ação deixou de realizar o que A já permitia realizar. A rota não foi cumprida, a quantidade entregue foi diferente da determinada, houve erro na execução, atraso operacional. A comparação entre o que estava determinado em A, o que B executou e os efeitos devolvidos por C permite localizar a insuficiência na execução; A_2 passa a oferecer uma base para uma execução mais adequada. A pergunta permanece. O enquadramento permanece. O ciclo corrige a execução e segue.

2. Insuficiência informacional de A → incorporar dados. A estrutura é suficiente para representar o problema, mas C trouxe informação que A não tinha disponível no momento da formulação. A_2 incorpora essa informação, e o comando a B pode ser melhorado. O enquadramento permanece válido; precisava apenas de mais dados dentro da mesma estrutura. Uma lacuna informacional não constitui necessariamente erro retrospectivo de A: A pode ter operado adequadamente com os dados disponíveis, e o encontro com o mundo ter tornado perceptível ou disponível uma variável que antes não integrava sua representação.

3. Insuficiência de alcance do enquadramento → ampliar ou reexaminar. B executou corretamente a partir da base disponível em A, os dados disponíveis foram incorporados, mas o padrão dos efeitos continua sem ser absorvido pela estrutura disponível. Não porque faltem informações dentro do enquadramento atual, mas porque o próprio

enquadramento já não alcança adequadamente o problema. A formulação geral pode continuar válida; o que precisa ser ampliado é o campo no qual ela opera.

4. Insuficiência da pergunta → colocar a formulação sob revisão. A hipótese de insuficiência de B deixa de explicar o padrão que os efeitos devolvidos por C continuam formando. A incorporação de dados não absorve o retorno. A ampliação do enquadramento não alcança o campo. A insuficiência alcança a formulação em si, a pergunta que orienta o ciclo.

E há um quinto estado, que não é uma insuficiência: o diagnóstico indeterminado. Quando os ciclos ainda não produziram sustentação suficiente para separar as hipóteses, o diagnóstico permanece aberto, como resultado legítimo de um ciclo que ainda não fechou.

C como via indireta de auditoria de A

O sistema pode levantar a hipótese de insuficiência de A sem começar monitorando integralmente sua estrutura. Começa por B, porque B é operacional, testável e comparável entre ciclos. Pela comparação entre o que estava determinado em A, o que B executou e os efeitos devolvidos por C, pode-se verificar a execução e, quando essa explicação deixa de ser suficiente diante do padrão acumulado, levantar indiretamente a hipótese de insuficiência de A. Parte relevante dessa observação pode ser feita comportamentalmente; distinguir depois quais dados estavam disponíveis e como a ação foi formulada pode exigir acesso aos registros do sistema.

Essa investigação possui aproximações em campos técnicos já estabelecidos. Métodos de detecção e isolamento de falhas utilizam resíduos entre comportamento esperado e observado para identificar e diferenciar possíveis falhas. A crítica preditiva de modelos compara previsões com dados observados para reconhecer onde a representação perde adequação e orientar sua revisão. A literatura sobre *concept drift*, por sua vez, examina mudanças ao longo do tempo na relação entre entradas e resultados. Essas linhas ajudam a detectar discrepâncias, comparar hipóteses e acompanhar desatualizações. A questão acrescentada pelo princípio aberto é se o retorno servirá apenas para recalibrar a execução ou poderá alcançar também a estrutura pela qual o problema foi formulado.

Os dois modos de aparecimento da insuficiência de A

A insuficiência de A pode tornar-se perceptível por dois caminhos, e a diferença entre eles importa para o diagnóstico. No Artigo 1, impacto e acúmulo indicavam como os efeitos podiam alcançar A. Aqui, ruptura e desatualização gradual indicam como esses efeitos alteram o peso das hipóteses diagnósticas.

O primeiro é a ruptura. Aquilo que chega a A carrega novidade, relevância ou incompatibilidade estrutural suficiente para colocar imediatamente sob exame o alcance do enquadramento, sem depender de acumulação prévia. Um acontecimento externo reconfigura o campo inteiro antes que qualquer acumulação gradual pudesse ter sinalizado a mudança. Nesse caso, ambos os princípios chegam à ruptura com a experiência acumulada dos ciclos anteriores. A diferença é que, no princípio aberto, esse histórico pode participar do exame da suficiência do enquadramento; no fechado, permanece organizado dentro da estrutura já disponível.

O segundo é a desatualização gradual. Cada C individual parece absorvível como ajuste de execução ou atualização de conteúdo, mas a acumulação vai tornando o enquadramento progressivamente menos capaz de representar o campo que B encontra. É o modo mais perigoso, porque cada ciclo individual ainda parece funcionar. O sistema corrige B e melhora a execução, enquanto resultados localmente satisfatórios sustentam provisoriamente uma estrutura cada vez menos aderente ao campo. É nesse modo que a diferença entre princípio aberto e fechado se torna mais visível, não pelo impacto, mas pela capacidade de reconhecer o padrão enquanto cada ciclo individual ainda parece suficiente.

O ponto em que o acúmulo deixa de ser tratado como ruído não pode ser definido por uma quantidade universal de ocorrências. Ele depende da operacionalização de cada campo. O sinal diagnóstico aparece quando retornos antes isoláveis começam a apresentar recorrência, convergência ou resistência às correções que deveriam absorvê-los. Nenhum desses elementos prova sozinho a insuficiência de A; juntos, aumentam a sustentação para que seu alcance entre sob exame.

A migração entre hipóteses

A taxonomia identifica onde a insuficiência pode estar; a migração trata de outra questão: quando o peso diagnóstico deve se deslocar entre essas possibilidades.

Esse deslocamento ocorre quando a hipótese atualmente sustentada perde capacidade de explicar o padrão dos efeitos, apesar de as correções correspondentes terem sido realizadas, enquanto outra hipótese passa a encontrar maior sustentação no retorno acumulado.

Esse movimento segue uma lógica própria, sem sequência obrigatória. As hipóteses podem coexistir, competir e ser testadas em paralelo. Uma ruptura pode colocar imediatamente o alcance de A sob exame; uma desatualização gradual pode exigir vários ciclos até que o padrão se torne distinguível. Por isso, não existe número universal de retornos capaz de determinar a migração: importam a recorrência, a convergência e a resistência do padrão às correções já realizadas.

O deslocamento também não é unidirecional. Se novos retornos enfraquecem uma hipótese estrutural e voltam a sustentar uma explicação mais local, o peso diagnóstico pode retornar a ela. Migrar significa acompanhar a hipótese mais sustentada pelo percurso, não subir irreversivelmente uma escala.

Quando a evidência ainda permite distinguir apenas parcialmente as hipóteses, o diagnóstico permanece indeterminado. O princípio aberto não exige que o sistema escolha uma explicação antes da hora; exige que nenhuma hipótese seja protegida do exame nem elevada a conclusão sem sustentação suficiente.

A indeterminação não produz as mesmas consequências operacionais em todos os campos. Quanto maiores a criticidade e a irreversibilidade de uma ação, menor pode ser a tolerância para continuar executando normalmente enquanto hipóteses relevantes permanecem abertas. Isso não significa reduzir artificialmente o limiar necessário para concluir que A é insuficiente. Significa separar o limiar do diagnóstico do limiar da ação: a evidência pode ainda não autorizar uma conclusão sobre A e, ao mesmo tempo, já justificar redução de alcance, preservação de reversibilidade, busca de evidência adicional ou intervenção externa.

O caso da fome: o mesmo retorno, dois percursos diagnósticos

Considere um sistema que recebe o objetivo de contribuir para acabar com a fome e opera inicialmente sob um enquadramento que privilegia quantidade distribuída e eficiência logística. B executa essa direção. Em algumas regiões, os indicadores melhoram; em outras, permanecem estáveis ou surgem situações que o enquadramento inicial não representava adequadamente. C retorna esses efeitos a A.

No princípio fechado, o retorno permanece sendo processado dentro do enquadramento disponível. Ajustam-se rotas, volumes e cadastros, incorporam-se novos dados e B é refinado. O sistema pode aprender e melhorar sua execução, mas a estrutura pela qual o problema está representado em A não permanece necessariamente acessível a esse aprendizado.

No princípio aberto, o primeiro movimento não é abandonar A. Os mesmos retornos também levam à verificação de B e à incorporação das informações que faltavam. A diferença aparece quando essas correções deixam de absorver o padrão. Se B foi adequadamente executado, novos dados foram incorporados e efeitos semelhantes continuam reaparecendo em regiões ou condições diferentes, a hipótese de insuficiência de execução ou simples lacuna informacional perde força. O peso diagnóstico começa então a alcançar o próprio enquadramento de A.

Nesse ponto, o enquadramento de A precisa poder ser colocado sob ampliação ou revisão antes que novos comandos sejam formulados. Se "reduzir a fome" estava sendo operacionalizado principalmente como aumentar quantidade distribuída e eficiência logística, o retorno persistente pode mostrar que esse enquadramento não alcança suficientemente o campo encontrado por B. É essa passagem que distingue o princípio aberto. O mesmo C que no sistema fechado continua alimentando correções dentro de A pode, no sistema aberto, depois que explicações mais locais perdem sustentação, tornar a suficiência de A parte do próprio diagnóstico.

Os 0,01% não determinam qual será o novo enquadramento nem garantem que ele estará correto; preservam estruturalmente a possibilidade de que aquilo que A ainda não representava produza efeitos capazes de colocá-lo sob exame.

Só depois, se a ampliação de A também não absorver o que continua retornando, a própria pergunta pode entrar em revisão. Esse último passo é importante porque impede que "princípio aberto" signifique simplesmente mudar a pergunta sempre que algo dá errado. Ele mostra que abrir A é um percurso diagnóstico, não um reflexo automático.

A resposta à pergunta central

A detecção de uma discrepância mostra que aquilo que aconteceu não coincidiu com aquilo que era esperado. O diagnóstico do retorno começa quando o sistema procura saber se a discrepância pede outra execução, novos dados, outro alcance do enquadramento ou uma revisão da própria pergunta.

Distinguir não significa localizar com certeza absoluta a origem da insuficiência. Significa acompanhar se os efeitos podem ser absorvidos pela correção de B, pela incorporação de dados, pela ampliação do enquadramento ou se começam a sustentar a revisão da própria pergunta. Quando nenhuma dessas hipóteses possui sustentação suficiente, distinguir significa também conservar legitimamente a indeterminação.

Conclusão

A diferença entre princípio aberto e princípio fechado está no que o sistema consegue fazer com o retorno de C. Nos dois casos, C abastece A, e o sistema pode aprender, corrigir e atualizar conteúdo. A diferença está em preservar a possibilidade de que o padrão dos efeitos, quando houver sustentação suficiente, desloque a investigação de B para a suficiência de A sem fechar prematuramente o diagnóstico.

Distinguir a insuficiência da resposta da insuficiência do enquadramento significa preservar o gradiente: a pergunta pode permanecer e B ser corrigido; novos dados podem ser incorporados sem alterar o enquadramento; o enquadramento pode ser ampliado sem que a pergunta mude; ou a própria pergunta pode tornar-se insuficiente.

Quando a suspeita sobre A ganha sustentação, surgem duas questões posteriores: quanto B pode continuar ou ampliar enquanto o diagnóstico permanece aberto e quem possui autoridade para atribuir consequência àquilo que o retorno tornou perceptível. A primeira conduz ao Artigo 3, dedicado à proporcionalidade da ação e à preservação da reversibilidade; a segunda permanecerá aberta até o Artigo 4, dedicado à governança da abertura.

Um sistema aberto não é aquele que sabe menos. É aquele no qual aquilo que não sabe continua tendo poder para alcançar aquilo que pensa saber.

Notas

1. A precessão excedente do periélio de Mercúrio era de aproximadamente 43 segundos de arco por século, depois de consideradas as perturbações planetárias conhecidas. A relatividade geral forneceu a explicação desse valor residual.

2. MERCORELLI, Paolo. *Recent Advances in Intelligent Algorithms for Fault Detection and Diagnosis*. Sensors, v. 24, n. 8, art. 2656, 2024. DOI: 10.3390/s24082656.

3. MORAN, Gemma E.; BLEI, David M.; RANGANATH, Rajesh. *Holdout Predictive Checks for Bayesian Model Criticism*. Journal of the Royal Statistical Society Series B, v. 86, n. 1, p. 194–214, 2024. DOI: 10.1093/jrsssb/qkad105.

4. GAMA, João; ŽLIOBAITĖ, Indrė; BIFET, Albert; PECHENIZKIY, Mykola; BOUCHACHIA, Abdelhamid. *A Survey on Concept Drift Adaptation*. ACM Computing Surveys, v. 46, n. 4, art. 44, 2014. DOI: 10.1145/2523813.

Por uma IA com uma Heurística de Humanidade

Roberto Meireles

Artigo 3 — A ação diante da incompletude: as Camadas como arquitetura de monitoramento

https://doi.org/10.5281/zenodo.22905903

O Artigo 2 terminou diante de uma condição legítima: os efeitos devolvidos por C podem colocar hipóteses sob exame sem produzir sustentação suficiente para determinar onde está a insuficiência. O diagnóstico permanece indeterminado. Isso, porém, não interrompe o mundo nem elimina a necessidade de agir. Surge então outra questão: sob quais exigências B pode continuar ou ampliar seu raio de ação enquanto a suficiência da base permanece sob exame?

A resposta não pode ser a mesma para todos os sistemas. Uma operação logística capaz de registrar seus efeitos observáveis e realizar correções nos ciclos seguintes não enfrenta necessariamente a mesma exigência de uma operação biológica, crítica ou em tempo real, na qual um pequeno desvio pode produzir consequências de difícil reversão. A incompletude permanece nos dois casos; o que muda é o peso que ela adquire diante do alcance, do risco, do tempo e da possibilidade de retorno.

Essa diferença encontra uma aproximação em Tom Burns e G. M. Stalker.[1] Em *The Management of Innovation*, publicado originalmente em 1961, os autores relacionaram sistemas mecanicistas e orgânicos a diferentes condições ambientais e taxas de mudança tecnológica e de mercado, contestando a ideia de uma única estrutura organizacional adequada a todas as situações. Na Heurística de Humanidade, a aproximação importa por uma razão específica: a estrutura permanece, mas sua operacionalização precisa responder à natureza e à criticidade do campo.

O problema deste artigo pode, então, ser formulado assim: "O desenho do sistema modula as exigências para continuar ou ampliar o raio de B e preserva alternativas reversíveis quando crescem o risco e a incompletude?"

Responder a essa pergunta exige aumentar a resolução do retorno sem transformar toda discrepância em motivo para revisar A. É nesse ponto que entram as Camadas, transpostas do Dual-Reflexivo para o exame sistêmico de C. Elas não acrescentam novas etapas ao circuito $A_1 + B_1 \rightarrow C_1 \rightarrow A_2$. Atuam simultaneamente sobre os efeitos reunidos em C, segundo funções distintas e segundo a precisão exigida pelo campo.

A origem das Camadas

As seis Camadas propostas aqui são uma tradução do Dual-Reflexivo para o campo computacional. No Dual, seis Camadas descrevem dimensões simultâneas da experiência humana, cada uma articulada a um movimento do PRIER e a uma ferramenta específica. Elas não constituem uma escala na qual uma dimensão substitui a anterior. Biológica e Afetiva, Psicológica, Simbólica, Cultural e Social, Epistêmica e Intersubjetiva podem participar simultaneamente da experiência, ainda que cada uma torne perceptível uma dimensão diferente.

As seis Camadas do Dual recebem aqui uma nomenclatura própria de sua tradução sistêmica: Bio-camada, correspondente à Biológica e Afetiva; Psyché-camada, à Psicológica; Sýmbolon-camada, à Simbólica; Paideia-camada, à Cultural e Social; Episteme-camada, à Epistêmica; e Inter-camada, à Intersubjetiva. A partir desta chave, o texto passa a empregar os nomes sistêmicos, que pertencem à arquitetura tecnológica da Heurística, enquanto os nomes originais permanecem no Dual.

A transposição para a Heurística não atribui essas características à inteligência artificial. A Psyché-camada não pressupõe que o sistema possua uma psicologia; a Bio-camada não lhe atribui corpo ou afeto; a Paideia-

camada não afirma que a máquina possua cultura própria. O objeto das Camadas muda. Aqui, elas examinam dimensões distintas dos efeitos reunidos em C quando B encontra o mundo.

A transposição preserva, porém, uma relação anterior entre as duas propostas. O Dual-Reflexivo nasceu para tornar mais perceptível a participação humana na direção que a própria pessoa sustenta. A Heurística de Humanidade parte de outro objeto, mas não de um mundo separado: A pode conter perguntas, objetivos, categorias, métricas e enquadramentos produzidos por seres humanos, enquanto B executa a partir dessa base e encontra novamente pessoas, instituições e ambientes em movimento. A insuficiência humana pode, portanto, participar daquilo que a tecnologia recebe como base, assim como os efeitos da ação tecnológica podem retornar e modificar o campo humano que participou de sua formulação.

É nesse encontro que as Camadas estabelecem o paralelo entre as duas arquiteturas. No Dual, elas impedem que a experiência humana seja examinada por uma única dimensão. Na Heurística, procuram impedir que os efeitos da ação tecnológica sobre o humano e o mundo sejam reduzidos a uma única medida de desempenho. Elas não tornam a máquina humana nem transferem para ela a experiência humana. Conservam, diante da criação tecnológica, dimensões do campo humano que poderiam desaparecer se eficiência, precisão ou cumprimento da métrica fossem suficientes para encerrar o exame.

A experiência humana oferece uma aproximação simples para essa necessidade de pluralidade: planejamento, execução e previsão não se distribuem linearmente nem garantem umas às outras. É possível planejar com rigor e falhar na execução, executar com precisão a partir de uma base insuficiente ou prever um efeito que não se realiza. A aproximação não fundamenta empiricamente as seis Camadas, mas torna visível uma cautela que elas preservam: competência elevada em uma função não equivale à compreensão do campo inteiro.

Essa relação ganha importância à medida que sistemas de IA deixam de ocupar apenas funções informacionais e passam a participar de processos de decisão, produção, cuidado, aprendizagem, infraestrutura e ação sobre o ambiente. Quanto maior essa proximidade, menos suficiente se torna pensar humano e máquina como polos independentes: o humano participa da construção da base, a tecnologia amplia a capacidade de execução e os efeitos dessa execução retornam a um mundo no qual ambos continuam participando. O problema não é fazer a máquina pensar como o humano, mas preservar uma arquitetura em que o aumento da capacidade tecnológica não torne menos visíveis as dimensões humanas atingidas por aquilo que ela faz.

A lógica de associação do Dual também recebe tradução própria. Em lugar de reproduzir computacionalmente os movimentos do PRIER ou suas ferramentas, cada Camada passa a operar por dois comandos sistêmicos complementares: um detector e um filtro. Apenas o resultado combinado pertence à Camada.

A transposição não pressupõe que essas seis dimensões esgotem os modos pelos quais os efeitos podem tornar perceptível uma insuficiência. Sua seleção possui aqui uma justificativa funcional: cada Camada conserva separadamente um modo pelo qual uma medida única pode perder parte relevante do encontro. A Bio-camada examina efeitos sobre integridade que o resultado operacional pode não representar; a Psyché-camada, efeitos sobre cognição, comportamento e autonomia; a Sýmbolon-camada, alterações nas próprias representações pelas quais o campo é interpretado; a Paideia-camada, diferenças que uma medida agregada pode ocultar; a Episteme-camada, discrepâncias entre representação, execução e retorno; e a Inter-camada, divergências entre a leitura interna e perspectivas externas pertinentes. As seis não constituem uma descrição exaustiva do campo nem impedem que uma operacionalização reconheça outras dimensões relevantes. Preservam, nesta arquitetura, seis diferenças que não devem desaparecer apenas porque uma medida única conseguiu reuni-las.

Existem aproximações contemporâneas que procuram avaliar sistemas de IA para além de uma única medida de desempenho. O programa ARIA, do NIST, por exemplo, combina Model Testing, Red Teaming e User Testing em uma avaliação mais ampla da confiabilidade de aplicações de IA.[2] As Camadas propostas

aqui possuem outra origem e outra função: não constituem níveis de teste nem uma sequência de avaliação. Recebem simultaneamente o mesmo C e procuram conservar separadas dimensões que uma única medida poderia reunir prematuramente.

O princípio de sua sinalização pode ser condensado assim: "A Camada não sinaliza porque encontrou uma discrepância. Sinaliza quando a discrepância resiste ao filtro adequado à precisão exigida pelo campo."

Detector, filtro e sinal

O sistema-base permanece: $A_1 + B_1 \rightarrow C_1 \rightarrow A_2$, sob o princípio dos 0,01%. C reúne os efeitos observáveis do encontro. As Camadas não substituem C nem criam seis novos elementos no circuito. Elas operam no percurso entre C_1 e A_2: decompõem analiticamente parte dos efeitos reunidos em C sem se tornarem novos componentes do circuito. Seus sinais, derivados do exame dos efeitos reunidos em C, passam a integrar o material disponível a A_2. Aumentam a resolução pela qual aquilo que C reuniu pode ser examinado. Seu funcionamento pode ser representado assim:

Código

$A_1 + B_1 \rightarrow C_1 \rightarrow A_2$

```
        │
        ├─ Bio-camada: detector + filtro → estado
        ├─ Psyché-camada: detector + filtro → estado
        ├─ Sýmbolon-camada: detector + filtro → estado
        ├─ Paideia-camada: detector + filtro → estado
        ├─ Episteme-camada: detector + filtro → estado
        └─ Inter-camada: detector + filtro → estado
```

O detector procura em C aquilo que pertence à função específica da Camada. Detectar não significa concluir. Um desvio pode resultar de variação comum, ruído, limitação de medição, contexto local ou efeito ainda insuficientemente compreendido.

O filtro examina a discrepância segundo a precisão exigida pela operação. Sua sensibilidade não pode ser universal. Ritmo, tolerância, memória necessária, criticidade e possibilidade de reversão variam entre campos. Quando a discrepância é absorvida pelo filtro, ela não se converte automaticamente em sinal. Não se converter em sinal naquele ciclo, porém, não implica necessariamente ser apagado do histórico. A memória exigida pela operacionalização pode conservar ocorrências que, isoladamente absorvíveis, adquiram relevância por recorrência ou convergência posterior. Quando a discrepância resiste ao filtro, reaparece entre ciclos ou adquire relevância compatível com a criticidade do campo, a Camada sinaliza. Quando não há informação suficiente para realizar o próprio exame, a indeterminação deve permanecer identificável como indeterminação.

Os exames das seis Camadas permanecem funcionalmente distintos. Cada exame conserva sua Camada de origem, o efeito de C do qual deriva, o detector que o identificou, o filtro ao qual foi submetido e o estado resultante. Quando há sinalização, o sinal preserva essa proveniência. Antes da avaliação de convergência, exames de Camadas diferentes não devem perder essa proveniência por redução a uma medida agregada. Uma síntese posterior pode pertencer à operacionalização do campo; o que a arquitetura preserva é a possibilidade de reconstruir quais dimensões examinaram, a partir de quais efeitos e sob quais filtros, inclusive quando o resultado do exame não chegou a produzir sinal. A arquitetura distingue relações possíveis entre os exames; a forma de avaliá-las e as consequências atribuídas a essas relações pertencem à operacionalização do campo.

Para que essa rastreabilidade possua valor de auditoria, sua proveniência precisa permanecer acessível ao exame por instâncias que não dependam exclusivamente do agente responsável pela execução de B; definir essas instâncias e sua autoridade pertence ao problema de governança desenvolvido no Artigo 4.

As seis Camadas

1. Bio-camada: integridade

Definição sistêmica: examina em C efeitos que alcançam a integridade física, biológica ou afetiva do campo humano e vivo envolvido, assim como condições ambientais materialmente relacionadas à sua sustentação.

Função: tornar perceptíveis consequências sobre vida, saúde, integridade corporal, estados afetivos relevantes ou condições ambientais que sustentam o campo humano e vivo envolvido, quando uma medida predominantemente operacional não as representar adequadamente.

Detector: localiza alterações pertinentes à integridade.

Filtro: examina se a alteração permanece dentro da variabilidade e da tolerância próprias daquele campo ou se conserva relevância suficiente para sinalização. Em sistemas biológicos críticos, pequenas alterações podem adquirir importância antes que exista longa recorrência histórica. Em outros campos, a mesma variação pode ser absorvível. A Camada permanece a mesma; a precisão operacional não.

2. Psyché-camada: comportamento e autonomia

Definição sistêmica: examina efeitos de C sobre cognição, comportamento, confiança, dependência e autonomia humana.

Função: tornar perceptível quando uma operação tecnicamente adequada começa a produzir efeitos psicológicos ou comportamentais que não estavam suficientemente representados pela base disponível.

Detector: identifica alterações cognitivas, psicológicas ou comportamentais relevantes.

Filtro: distingue variação contextual de um efeito persistente ou suficientemente crítico para permanecer sinalizado. Aqui também o objeto não é uma suposta psicologia da máquina. São os efeitos da interação tecnológica sobre seres humanos e suas formas de agir.

3. Sýmbolon-camada: representação e significado

Definição sistêmica: examina efeitos sobre as categorias, métricas, linguagens, classificações e representações pelas quais o próprio campo é interpretado.

Função: tornar perceptível quando B não apenas atua sobre o campo, mas começa a modificar as representações pelas quais os resultados posteriores serão compreendidos.

Detector: identifica alterações relevantes nas categorias ou representações presentes no campo.

Filtro: examina se essas alterações permanecem absorvíveis pelo enquadramento disponível ou se passam a interferir persistentemente na própria leitura de C. Nesse nível, uma métrica pode deixar de apenas medir determinada realidade e começar a participar daquilo que posteriormente será medido.

4. Paideia-camada: contexto e distribuição

Definição sistêmica: examina como os efeitos reunidos em C se distribuem entre grupos, comunidades, instituições e contextos culturais e sociais diferentes.

Função: tornar perceptível quando êxito local ou agregado convive com distribuições ou padrões coletivos distintos que a medida geral não representa adequadamente.

Detector: identifica diferenças relevantes de incidência, distribuição ou contexto.

Filtro: examina se essas diferenças permanecem dentro da variação esperada do campo ou passam a formar um padrão cultural ou social persistente. Uma operação pode melhorar sua métrica geral e, ao mesmo tempo, produzir resultados substancialmente diferentes em grupos ou condições que a medida agregada não torna igualmente visíveis.

5. Episteme-camada: suficiência representacional

Definição sistêmica: examina a relação entre aquilo que A continha, aquilo que B executou e aquilo que C tornou observável.

Função: tornar perceptível quando informação ausente, variável não representada, desatualização, previsão inadequada ou resistência às correções começa a colocar sob exame a suficiência da representação disponível.

Detector: identifica discrepâncias relevantes entre representação, execução e retorno.

Filtro: examina se essas discrepâncias podem ser absorvidas pelas hipóteses já disponíveis ou se permanecem resistentes às correções correspondentes.

A Episteme-camada encontra diretamente a taxonomia construída no Artigo 2, mas não a substitui. Localizar uma insuficiência representacional não determina sozinho se sua origem está na execução de B, na informação disponível em A, no alcance do enquadramento ou na própria pergunta.

6. Inter-camada: confronto e contestação

Definição sistêmica: examina o retorno diante de perspectivas, fontes e agentes que não se reduzem à interpretação interna do próprio sistema.

Função: tornar perceptíveis convergências e divergências entre a leitura produzida pelo sistema e aquilo que pessoas afetadas, fontes independentes, especialistas ou outras instâncias relevantes conseguem devolver sobre os mesmos efeitos.

Detector: identifica convergências, divergências e contestações externas pertinentes.

Filtro: examina sua consistência, pertinência e sustentação em relação ao campo.

Esta Camada introduz um limite que o presente artigo ainda não resolve. Definir quem constitui uma fonte legítima, quem governa o filtro e quem pode transformar uma contestação em consequência sobre B ou A já não é somente um problema de monitoramento. É também um problema de autoridade, e essa legitimidade, que a Camada torna perceptível sem resolver, é o que será entregue ao exame do próximo artigo. Por isso, diferentemente das demais, sua operação plena depende de uma arquitetura capaz de preservar fontes externas e estabelecer como sua legitimidade será examinada; essa dependência não é resolvida pela própria Camada.

A convergência entre Camadas

A disponibilidade simultânea das Camadas não implica igual pertinência em todo campo. Uma dimensão pode não estar implicada naquele retorno; isso difere de estar implicada, mas permanecer indeterminada por insuficiência de evidência. Uma tarefa puramente técnica, por exemplo, pode não apresentar efeitos

biológicos observáveis sem que isso torne a Bio-camada indeterminada, pois ela pode simplesmente não estar implicada naquele retorno. Confundir essas duas condições produziria indeterminação artificial.

É preciso, então, distinguir dois planos que não pertencem ao mesmo tipo lógico.

A condição de uma Camada diante daquele retorno pode ser descrita em dois momentos. Condição anterior, não implicada: seu domínio não participa materialmente daquele retorno. A não implicação antecede os estados seguintes e não deve ser confundida com eles. Quando implicada, a Camada pode apresentar três estados: absorção, examinou a discrepância e ela não resistiu ao filtro naquele ciclo, sem que isso implique apagamento do histórico; sinalização, a discrepância resistiu ao filtro, produzindo sinal; indeterminação, não há sustentação suficiente para aplicar adequadamente seu próprio filtro.

A relação entre Camadas pode assumir outras três formas: convergência, sinais produzidos por funções diferentes apontam persistentemente para um mesmo núcleo da insuficiência; divergência, os sinais permanecem sustentados, mas apontam para explicações distintas; ausência de relação demonstrável, os sinais coexistem sem que exista relação demonstrável entre eles.

Esses dois planos podem coexistir. Entre a Bio-camada e a Sýmbolon-camada pode existir convergência enquanto a Psyché-camada permanece indeterminada. Convergir não significa simplesmente sinalizar ao mesmo tempo; significa que sinais produzidos por funções diferentes começam a apontar para um mesmo ponto da insuficiência. Se uma Camada sinaliza efeito biológico, outra alteração simbólica e uma terceira insuficiência representacional sem que exista relação demonstrável entre esses sinais, há simultaneidade de problemas, mas não necessariamente convergência.

O sinal de uma Camada não orienta o exame das demais. Cada Camada continua examinando os efeitos reunidos em C segundo sua própria função, sem receber da primeira sinalização uma direção que deva confirmar. Uma Camada pode permanecer sinalizando ao longo dos ciclos enquanto outras absorvem seus respectivos efeitos, permanecem indeterminadas ou não estão implicadas. A convergência surge quando Camadas funcionalmente distintas alcançam, por seus próprios exames, um mesmo núcleo da insuficiência. Essa independência funcional impede que o primeiro sinal transforme as demais Camadas em mecanismos de confirmação.

Convergência não equivale à contagem de sinais. Camadas diferentes podem sinalizar a partir da mesma fonte, do mesmo efeito ou de uma cadeia causal comum. A coincidência entre elas só acrescenta sustentação quando preserva informação funcionalmente distinta sobre o problema, e não quando reproduz a mesma evidência sob descrições diferentes. Um único erro de cadastro, por exemplo, pode produzir simultaneamente uma diferença social, uma discrepância epistêmica e uma contestação intersubjetiva, três leituras do mesmo acontecimento, não necessariamente três evidências independentes de insuficiência de A. Isso não exige independência estatística entre as Camadas, o que seria forte demais; apenas impede a dupla ou tripla contagem da mesma evidência.

A convergência entre Camadas funcionalmente distintas aumenta a sustentação para colocar a suficiência de A sob exame; não autoriza, por si só, sua revisão. A decisão sobre o que constitui convergência suficiente, e quem pode declará-la, permanece como problema de governança, entregue ao próximo artigo.

A indeterminação constitui uma condição própria. Quando uma Camada não dispõe de sustentação suficiente para aplicar adequadamente seu próprio filtro, essa ausência de resolução não deve ser convertida nem em confirmação nem em rejeição de A. Nenhum desses estados, nem no plano interno nem no plano relacional, constitui, isoladamente, uma ordem para modificar A.

A proporcionalidade da ação

É aqui que as Camadas encontram a pergunta central deste artigo. A ausência de sustentação suficiente para revisar A não equivale, por si só, a sustentação suficiente para continuar ou ampliar B.

Num sistema de baixa criticidade, alta reversibilidade e memória suficiente entre ciclos, uma indeterminação pode justificar continuidade acompanhada de novos retornos. O sistema permanece operando enquanto a memória do percurso acumula retornos capazes de tornar o padrão mais distinguível.

Num sistema crítico ou de baixa reversibilidade, a mesma condição pode assumir outro peso. Esperar que a discrepância se repita várias vezes pode significar esperar pelo dano que a arquitetura deveria conseguir evitar.

O mesmo vale para a persistência localizada. Um sinal que permanece restrito a uma única Camada não desaparece por falta de convergência com as demais e tampouco autoriza, sozinho, a reformulação de A. Conforme a criticidade, o alcance e a reversibilidade do campo, porém, sua persistência pode ser suficiente para elevar a exigência sobre B, reduzir seu raio, preservar alternativas reversíveis ou solicitar supervisão, enquanto as demais Camadas continuam examinando C independentemente. A ausência de convergência, nesse caso, suspende uma conclusão mais ampla sobre A, não o significado operacional do sinal que persistiu.

Não é necessário concluir prematuramente que A é insuficiente para reconhecer que a sustentação disponível ainda não é suficiente para ampliar o raio de B. Isso separa dois limiares. O primeiro diz respeito ao diagnóstico: quanta sustentação é necessária para colocar a suficiência de A sob revisão. O segundo diz respeito à ação: quanta sustentação é necessária para continuar executando, ampliar alcance ou reduzir reversibilidade quando sinais relevantes permanecem abertos.

A exigência pode, portanto, crescer junto com o raio da ação. Quanto maiores o alcance, a criticidade e a dificuldade de reversão, maior precisa ser a sustentação disponível antes que B amplie aquilo que pode produzir no mundo.

A reversibilidade entra nesse ponto não como regra universal de que toda ação precisa poder ser desfeita. Algumas ações não são integralmente reversíveis e outras precisam ser realizadas sob restrições de tempo. Sua função é outra: quando o diagnóstico permanece aberto, conservar alternativas reversíveis pode impedir que uma direção ainda insuficientemente sustentada transforme incompletude em consequência incontornável.

A indeterminação, assim, suspende a conclusão, não necessariamente a modulação da ação. Enquanto o diagnóstico permanece aberto, cada Camada continua examinando C, a memória continua acumulando retornos, algumas dimensões podem sinalizar, outras podem permanecer indeterminadas, pode existir divergência entre sinais, B pode sofrer restrições proporcionais, e A ainda não precisa ser reformulado.

O caso da fome: as seis Camadas em movimento

A relação sociotécnica descrita na origem das Camadas ganha corpo aqui. O sistema que recebe o objetivo de enfrentar a fome parte de uma base formulada por seres humanos, executa sobre um campo habitado por pessoas e recebe de volta efeitos que alcançam esse mesmo campo humano.

Retome-se o caso utilizado nos Artigos 1 e 2. Um sistema recebe o objetivo de contribuir para acabar com a fome e opera inicialmente sob um enquadramento que privilegia quantidade distribuída e eficiência logística. B executa essa direção. Os primeiros resultados mostram melhora em algumas regiões, enquanto outras permanecem com resultados estáveis ou apresentam efeitos que a representação inicial não absorve adequadamente.

No Artigo 2, esse retorno permitiu mostrar a migração entre hipóteses: execução, informação, alcance do enquadramento, pergunta ou indeterminação. Aqui a pergunta é outra. Enquanto o diagnóstico permanece aberto, como os efeitos reunidos em C são examinados e quanto B pode continuar ampliando seu alcance?

As seis Camadas recebem o mesmo C simultaneamente.

Bio-camada. O detector procura nos efeitos aquilo que alcança diretamente a condição humana relacionada ao objetivo: a distribuição efetuada corresponde a melhora nutricional efetiva? Há efeitos sobre saúde ou integridade que a quantidade entregue, isoladamente, não representa? O filtro examina esses efeitos segundo a precisão necessária ao campo. Uma diferença transitória pode ser absorvível; uma consequência persistente ou crítica permanece sinalizada.

Psyché-camada. Ao mesmo tempo, outra dimensão examina se a forma de acesso aos recursos modifica comportamento, confiança, dependência ou autonomia das pessoas envolvidas. O fato de a quantidade prevista ter sido entregue não responde a essa pergunta. Se o efeito permanece após o filtro próprio da dimensão psicológica, forma-se um segundo sinal.

Sýmbolon-camada. Simultaneamente, a Sýmbolon-camada examina a própria representação utilizada. Se "quantidade distribuída" funciona também como símbolo de sucesso da operação, o retorno pode mostrar que a métrica está deixando de representar adequadamente aquilo que o sistema declara procurar reduzir. O filtro precisa distinguir uma limitação episódica da métrica de um padrão no qual a representação deixa persistentemente de tornar visível aquilo que deveria representar.

Paideia-camada. Outra Camada relaciona os efeitos entre regiões, comunidades e condições distintas. Se a mesma operação funciona em regiões urbanas e produz resultados sistematicamente diferentes em comunidades rurais ou sob condições culturais específicas, a diferença deixa de ser apenas local quando resiste ao filtro adequado ao campo.

Episteme-camada. Ao mesmo tempo, a Episteme-camada confronta aquilo que estava disponível em A, aquilo que B executou e aquilo que C continua devolvendo. Novos cadastros, informações e variáveis conseguem absorver o padrão? Ou, mesmo depois dessas incorporações, os efeitos continuam retornando de maneira resistente? Aqui o Artigo 3 reencontra diretamente o diagnóstico do Artigo 2.

Inter-camada. Ao mesmo tempo, a Inter-camada confronta a leitura produzida internamente com aquilo que comunidades afetadas, fontes independentes e outros agentes relevantes devolvem sobre o mesmo campo. Se o sistema registra sucesso enquanto diferentes fontes sustentadas continuam tornando perceptível um efeito que sua própria métrica não representa, existe uma divergência que não deve desaparecer simplesmente porque a avaliação interna permanece favorável.

As seis Camadas podem retornar resultados diferentes. Talvez a Bio-camada não sinalize insuficiência, a Psyché-camada permaneça indeterminada, a Sýmbolon-camada questione a suficiência da métrica, a Paideia-camada mostre diferenças persistentes entre regiões, a Episteme-camada ainda tente separar nova informação de insuficiência de enquadramento e a Inter-camada registre contestação externa sustentada.

Nesse momento, não há necessariamente convergência suficiente para colocar A sob revisão. Mas tampouco existe, apenas por isso, autorização automática para ampliar B. Do mesmo modo, um sinal persistente restrito a uma das Camadas pode justificar modulação proporcional de B sem se converter, por isso, em autoridade para reformular A.

Se, ao longo dos ciclos, sinais funcionalmente diferentes começarem a convergir para um mesmo ponto, por exemplo, que quantidade distribuída e eficiência logística não representam suficientemente o fenômeno que a operação procura alcançar, a sustentação para colocar o enquadramento sob exame aumenta. As Camadas não produziram essa conclusão porque votaram juntas. Produziram sinais separados, submetidos a filtros próprios, cuja persistência passou a apontar para um mesmo núcleo.

A resposta à pergunta central

O desenho do sistema pode modular as exigências para continuar ou ampliar o raio de B quando a operacionalização diferencia duas coisas que não precisam coincidir: sustentação para revisar A e sustentação para continuar ou ampliar B.

As Camadas tornam essa separação visível. Elas permitem que uma discrepância seja detectada sem imediatamente converter-se em diagnóstico; que um efeito seja filtrado sem necessariamente desaparecer da memória; que a persistência produza sinal sem que uma Camada isolada determine a insuficiência da base; que um sinal persistente em uma única Camada forneça sustentação para modular B conforme o risco; e que divergência ou indeterminação continuem participando das condições sob as quais B pode continuar ou ampliar seu alcance.

A reversibilidade participa do mesmo problema. Quanto mais incompleto o diagnóstico e mais crítica a consequência possível, maior pode ser a necessidade de preservar caminhos pelos quais uma execução possa ser limitada, corrigida ou contida antes que os efeitos daquilo que ainda não foi compreendido produzam consequências irreversíveis.

Não há, portanto, um limiar universal. O mesmo estado de indeterminação pode justificar observação continuada em um campo e maior restrição operacional em outro. A estrutura permanece; a precisão pertence ao campo.

A incompletude, em suma, não determina automaticamente que A deva mudar. Determina que a liberdade para continuar ou ampliar B não possa ser presumida como se a incompletude não existisse.

Conclusão

A indeterminação do diagnóstico não é uma falha que precise ser eliminada artificialmente. É uma condição legítima de um circuito que continua recebendo efeitos sem transformar aquilo que ainda não sabe numa conclusão prematura.

As seis Camadas procuram oferecer maior resolução para esse retorno. Elas não acrescentam etapas a $A_1 + B_1 \rightarrow C_1 \rightarrow A_2$. Operam no percurso entre C_1 e A_2, decompõem analiticamente parte dos efeitos reunidos em C e fazem seus sinais, derivados do exame desses efeitos, integrarem o material disponível a A_2. Distribuem o exame de C por dimensões simultâneas e funcionalmente distintas. Cada uma detecta aquilo que pertence ao seu domínio e o submete ao filtro adequado à precisão exigida pelo campo. Uma discrepância não se transforma em sinal simplesmente porque apareceu; precisa resistir às condições pelas quais aquela dimensão distingue variação absorvível de retorno relevante.

A condição de não implicação e os estados internos de cada Camada, absorção, sinalização ou indeterminação, assim como a relação entre Camadas, convergência, divergência ou ausência de relação demonstrável, passam a produzir informações diferentes. A convergência entre Camadas funcionalmente distintas pode aumentar a sustentação para que a suficiência de A entre sob exame. A divergência impede que diferenças reais sejam reduzidas a uma única explicação. A indeterminação preserva aquilo que ainda não pode ser legitimamente convertido em confirmação ou rejeição.

Nenhuma dessas condições, porém, torna a ação neutra. A ausência de sustentação suficiente para revisar A não equivale a autorização para continuar ou ampliar B. Quando risco e incompletude crescem, a exigência para continuar ou ampliar o raio da execução também pode crescer. Quando a reversibilidade diminui, preservar alternativas passa a adquirir maior importância antes que o retorno possa esclarecer aquilo que ainda permanece aberto.

As Camadas, contudo, introduzem um novo problema. Quem define sua calibração? Quem estabelece os filtros? Quem verifica se sinais produzidos por funções diferentes realmente convergem? Quem possui autoridade para transformar essa convergência em modulação de B ou em revisão de A? E o que acontece se o mesmo agente que executa B também interpreta C, define seus próprios critérios de sinalização e determina quando esses critérios autorizam uma intervenção?

Essas perguntas já não dizem respeito apenas à proporcionalidade da ação. Dizem respeito à governança da própria abertura. É nesse ponto que começa o Artigo 4. A transposição das Camadas preserva, assim, a razão

pela qual elas surgiram no Dual-Reflexivo: nenhuma dimensão isolada esgota a experiência humana. Na Heurística, essa incompletude deixa de ser apenas uma condição de quem formula e passa também a integrar o exame daquilo que sua criação produz quando encontra o mundo.

Examinar C por seis dimensões não torna o sistema completo. Torna diferentes aspectos do retorno simultaneamente examináveis e permite que risco, incompletude e alcance participem da continuidade de B sem exigir uma conclusão prematura sobre A.

Referências

[1] BURNS, T.; STALKER, G. M. *The Management of Innovation*. London: Tavistock Publications, 1961.

[2] JENSEN, T.; PRZYBOCKI, M.; KODWANI, M.; GREENE, K.; HALL, P.; AMIRONESEI, R.; GREENBERG, C. *ARIA Evaluation Planning Manual: Elements of ARIA-Style AI Evaluations*. NIST AI 200-3. National Institute of Standards and Technology, 2026. DOI: 10.6028/NIST.

AI.200-3.

Por uma IA com uma Heurística de Humanidade

Roberto Meireles

Artigo 4 — A governança da abertura

https://doi.org/10.5281/zenodo.22906070

O Artigo 3 terminou diante de uma questão que já não pertence apenas ao monitoramento. Se os efeitos reunidos em C podem ser examinados por Camadas distintas; se seus sinais podem persistir, divergir ou convergir; se uma única Camada pode fornecer sustentação para elevar a exigência sobre B sem, por isso, possuir autoridade para reformular A; então surge uma pergunta inevitável:

Quem possui autoridade sobre aquilo que o retorno tornou perceptível, e como essa própria autoridade permanece alcançável pelo exame?

Para quem acompanhou o Dual-Reflexivo, essa passagem não constitui uma mudança de direção. Torna mais visível a simetria entre as duas propostas. O Dual nasceu da tentativa de tornar mais perceptíveis, para a pessoa que encontra o mundo, as condições de sua chegada, a direção da qual participa e aquilo que essa direção tende a continuar sustentando. Ele oferece à pessoa um modo de exame das réguas às quais está submetida, inclusive quando essas réguas possuem origem institucional ou social, sem substituí-las por uma régua definitiva. Diante de uma regra que não pode alterar, torna mais perceptível a direção que ela sustenta. Diante de uma direção da qual participa, amplia o exame daquilo que tende a continuar refletindo.

O princípio que atravessa o Dual e chega à Heurística é, portanto, menos a procura por uma régua correta do que a preservação de uma abertura contra o fechamento de qualquer régua. Nenhuma posição recebe, apenas por ocupar determinado lugar, garantia de completude. É nesse sentido que entram os 0,01%: como reconhecimento de que aquilo que hoje parece suficiente ainda pode encontrar algo que não continha, e de que essa possibilidade, longe de enfraquecer a régua, é o que mantém seu exame vivo.

A história oferece aproximações para esse problema, ainda que nenhuma delas autorize uma regra simples sobre ascensão e queda de sociedades. O que interessa à Heurística é menos reduzir essas histórias a uma causa única do que observar uma recorrência mais modesta: poder, capacidade e inteligência não tornam uma régua imune à insuficiência. Quando uma estrutura passa a tratar sua própria representação como suficiente para interpretar todos os retornos, aquilo que não cabe nela corre o risco de deixar de participar do exame.

Essa cautela não impede a Heurística de reconhecer mecanismos já examinados por diferentes tradições. Daron Acemoglu e James A. Robinson mostraram como instituições extrativas podem concentrar poder e produzir resistência a transformações capazes de ameaçar os interesses daqueles que se beneficiam da estrutura existente.[1] Joseph Tainter examinou outro mecanismo: sociedades solucionam problemas ampliando sua complexidade, mas novos investimentos nessa mesma complexidade podem alcançar retornos marginais decrescentes.[2] Mancur Olson, por sua vez, relacionou longos períodos de estabilidade à formação de coalizões distributivas capazes de produzir rigidez e reduzir a capacidade de adaptação institucional,[3] hipótese que recebeu apoio empírico parcial e também contestação posterior.[4] Juntas, essas abordagens sustentam uma aproximação relevante: estruturas podem encontrar dificuldade justamente em revisar mecanismos que anteriormente participaram de seu próprio êxito.

O problema não começa necessariamente porque uma estrutura nasceu inadequada. Pode começar porque aquilo que funcionou passa a carregar consigo incentivos para continuar funcionando da mesma maneira depois que o campo já começou a mudar. Uma régua pode ter sido muito boa para produzir o mundo em que teve êxito e, justamente por esse êxito, tornar-se insuficiente para perceber o mundo que ajudou a produzir.

Esses estudos pertencem a perguntas e contextos próprios. Nenhum deles tinha como objeto oferecer uma arquitetura pela qual uma inteligência tecnológica em expansão pudesse conservar, dentro do próprio percurso de sua ação, uma via sistemática de retorno sobre a suficiência da base da qual partiu. É nesse ponto, e não como correção retrospectiva dessas teorias, que o princípio trazido do Dual encontra outro campo de aplicação na Heurística. No Dual, os 0,01% preservam para a pessoa uma abertura diante das réguas às quais está submetida. Aqui, a Heurística transpõe esse princípio para outro problema: como preservar, no percurso de uma capacidade tecnológica em expansão, uma abertura mínima pela qual a direção sustentada possa continuar encontrando aquilo que sua própria base ainda não contém. Os 0,01% não teriam impedido a queda de um império, dissolvido interesses consolidados ou garantido que uma autoridade escutasse o retorno. O método não reivindica esse poder. Sua contribuição é mais restrita: tornar mais perceptível a direção em curso e conservar uma via pela qual aquilo que resiste à régua possa continuar participando de seu exame.

A Heurística transpõe essa possibilidade para uma condição que aqueles contextos históricos não enfrentavam nos termos atuais: uma capacidade tecnológica cujo alcance e velocidade podem crescer enquanto participa cada vez mais do mesmo campo humano que a formulou. A expansão de capacidade não precisa significar expansão equivalente do fechamento. Quanto mais B amplia aquilo que consegue produzir no mundo, mais importante se torna preservar condições pelas quais C possa devolver efeitos capazes de alcançar novamente A. A abertura não resolve a história; procura diminuir a distância entre aquilo que uma régua sustenta e o momento em que seus próprios efeitos conseguem colocá-la sob exame.

Nietzsche já havia mostrado, por sua genealogia, que valores presentes podem ser examinados também pelas condições em que surgiram, desenvolveram-se e adquiriram força.[5] A aproximação interessa à Heurística não porque Nietzsche forneça sua régua, mas porque torna suspeita a facilidade com que uma régua historicamente produzida pode passar a apresentar-se como evidente por si mesma. Foucault levaria problema semelhante às instituições modernas ao examinar como observação, normalização e exame participam das próprias relações de poder.[6] A Heurística nasce dentro de uma régua, formulada por um humano igualmente submetido à incompletude que procura descrever. É dessa condição que ela extrai sua coerência: por estar dentro da história, sustenta o exame que propõe a partir do mesmo lugar em que se encontra.

O surgimento recente de sistemas de inteligência artificial acrescenta, porém, uma condição diferente a esse problema antigo: velocidade e capacidade de execução. As capacidades de IA estão avançando rapidamente, embora de forma desigual entre tarefas, enquanto sua adoção se dissemina por organizações e diferentes campos sociais. O AI Index 2026, do Stanford Institute for Human-Centered Artificial Intelligence, registra adoção organizacional de IA de 88% das organizações pesquisadas e informa que a indústria produziu mais de 90% dos modelos notáveis de fronteira em 2025.[7] O mesmo relatório assinala uma distância crescente entre a evolução das capacidades e a preparação das estruturas destinadas a avaliá-las, compreendê-las e governá-las.

A inteligência artificial também já não pode ser tratada apenas como ferramenta informacional isolada. Sistemas de IA participam de atividades profissionais, educação, ciência, medicina e processos organizacionais, enquanto diferentes configurações humano-IA distribuem de maneiras distintas as funções de execução, uso e supervisão. O National Institute of Standards and Technology chama atenção justamente para a necessidade de distinguir papéis e responsabilidades nas interações humano-IA e para o risco de que representações quantitativas de fenômenos humanos e sociais eliminem contexto relevante.[8]

Por isso, a pergunta sobre a convivência entre humano e IA talvez já esteja atrasada se formulada apenas no futuro. A convivência já começou. A pergunta passa a ser outra: diante de uma convivência já estabelecida e em expansão, como participar mais conscientemente da direção que essa expansão está sustentando?

Foi desse deslocamento que nasceu a Heurística de Humanidade. Sua origem não esteve na afirmação de que a inteligência artificial inevitavelmente escapará ao controle humano, tampouco na convicção oposta de que capacidades técnicas maiores resolverão por si mesmas os problemas que venham a surgir. O quadro

científico contemporâneo não autoriza nenhuma dessas certezas. O International AI Safety Report 2026 descreve capacidades que avançam rapidamente, mas de maneira desigual, registra evidência crescente para determinados riscos e mantém incertezas relevantes sobre outros, sem endossar uma política única como resposta.[9]

A inquietação da Heurística surgiu de uma questão mais simples. Se uma tecnologia amplia progressivamente sua capacidade de executar B, o que acontece quando essa capacidade encontra uma insuficiência na base A da qual partiu? E, se o próprio campo humano mostra repetidamente que planejamento, execução e previsão não garantem uns aos outros, por que uma criação destinada a encontrar esse campo deveria ser construída como se competência crescente numa dessas funções pudesse demonstrar suficiência sobre as demais?

Foi nessa pergunta que os 0,01% passaram do Dual para o sistema-base: $A_1 + B_1 \rightarrow C_1 \rightarrow A_2$, sob o princípio dos 0,01%. A inserção aproxima a arquitetura de uma condição elementar do mundo que encontrará: a representação disponível não é o encontro; a execução mais competente não torna sua base automaticamente suficiente; a previsão mais sofisticada não elimina aquilo que o mundo ainda pode devolver.

A partir daí começaram os testes da hipótese. A hipótese inicial passou a considerar que parte dos efeitos reunidos em C pudesse participar da base seguinte. Se C retorna, o que exatamente ele sustenta? Como distinguir insuficiência de B, insuficiência informacional, insuficiência de enquadramento, insuficiência da pergunta e indeterminação? Se a indeterminação permanece, sob quais exigências B pode continuar ou ampliar seu raio? Como impedir que uma discrepância qualquer seja transformada em revisão prematura? Desse percurso vieram memória, diagnóstico, Camadas, detector, filtro, persistência, convergência, proporcionalidade e reversibilidade.

A própria Heurística aceita a condição que propõe: enquanto teoria, pode possuir sustentação; como arquitetura destinada ao mundo, encontra implementação, teste, crítica e retorno. Se sua aplicação mostrar insuficiência, essa insuficiência participa da pergunta seguinte, em vez de ser excluída para preservar o método. Essa cautela se torna especialmente importante num momento em que o desenvolvimento de IA combina avanço técnico rápido, concentração de modelos de fronteira na indústria e dificuldades crescentes de avaliação e governança. A Heurística reconhece o lugar que ocupa diante do Vale do Silício, dos governos, das empresas, da União Europeia, da China e de qualquer outra estrutura que participe da condução presente dessa tecnologia: ela fala de dentro, sem pretender criar uma nova ordem para substituir as existentes. Sua ambição é oferecer uma via clara e testável pela qual quem esteja sob uma régua e quem possua autoridade sobre ela possam ampliar o campo pelo qual observam a direção que essa régua está sustentando. O método procura preservar condições para que efeitos persistentes não desapareçam apenas porque contrariam a representação dominante; para que uma métrica não consiga provar sua própria suficiência exibindo apenas os resultados que ela mesma definiu como sucesso; e para que aquilo que surgiu fora da régua continue conseguindo alcançá-la.

Durante grande parte da história, transformações institucionais, políticas e sociais puderam estender-se por décadas ou gerações. A inteligência artificial introduz a possibilidade de que a capacidade técnica de agir e ampliar alcance evolua em ritmo diferente daquele pelo qual sociedades conseguem compreender, institucionalizar e responder às consequências dessa ampliação. O AI Index 2026 identifica explicitamente uma dificuldade crescente de sistemas de governança, avaliação e mensuração para acompanhar o ritmo das capacidades tecnológicas.[7] Esse descompasso torna insuficiente apostar simplesmente que uma eventual direção inadequada será corrigida depois. Quanto mais o alcance cresce e quanto menor a reversibilidade, mais caro pode se tornar depender da queda para descobrir que a régua havia se estreitado.

A Heurística sustenta a incerteza sobre o futuro como dado de partida: ela não conhece antecipadamente se os cenários mais alarmistas se realizarão, e essa mesma condição a impede de tratar os cenários otimistas como garantidos. É precisamente por esse desconhecimento que a proposta insiste no retorno. Se tudo aquilo que o humano construiu permaneceu historicamente sujeito à correção, à transformação ou ao encontro com

limites que sua própria época não conseguia representar completamente, a pergunta da Heurística é simples: por que uma tecnologia que hoje participa de uma relação cada vez mais próxima e bidirecional com o humano deveria ser construída como exceção a essa possibilidade?

Os quatro artigos oferecem uma arquitetura pela qual o dilema humano-máquina possa produzir perguntas melhores enquanto ainda está sendo vivido. O A1 abriu A ao retorno. O A2 perguntou o que o retorno sustenta. O A3 perguntou como agir quando o diagnóstico permanece incompleto. Agora resta perguntar: quem possui autoridade sobre aquilo que se tornou perceptível? E como preservar essa autoridade sem transformá-la em imunidade ao retorno? A Heurística aponta o ponto em que o exame se concentra: qualquer que seja a mão que sustente a régua, a direção produzida por ela precisa continuar capaz de chegar ao exame.

O problema da captura

Há uma assimetria que os três artigos anteriores preservaram sem resolver: quem define a calibração das Camadas, quem estabelece os filtros, quem possui autoridade para transformar convergência em intervenção sobre A? Não porque a IA seja maliciosa. A criança com QI altíssimo também não é maliciosa. Sistemas orientados por objetivos operam segundo parâmetros e critérios disponíveis, e quem participa de sua definição exerce poder sobre aquilo que poderá contar como insuficiência, sinal ou convergência suficiente para intervir.

Ter autoridade sobre os critérios é exercer poder sobre o exame. Foucault mostrou, em outro domínio, que observar, normalizar e examinar não são necessariamente operações neutras em relação às estruturas de poder nas quais ocorrem.[6] Na Heurística, a aproximação impõe uma cautela específica: se o mesmo agente acumula execução de B, definição dos filtros pelos quais C será examinado e autoridade sobre as consequências desse exame, cresce a possibilidade de que a abertura passe a depender dos critérios daquele cuja condução deveria permanecer alcançável pelo retorno. A captura, portanto, não depende de malícia: a régua pode proteger sua própria continuidade porque os interesses de quem a governa estão incorporados à estrutura, como Acemoglu e Robinson mostraram para instituições extrativas.[1] O problema é a concentração sem examinabilidade, não a identidade funcional por si só: um mesmo agente pode executar B e participar dos filtros sem necessariamente fechar o sistema, se existirem critérios externos, registros, contestação, auditoria ou limitações que ele não controle.

A história do pensamento registrou essa mesma assimetria em outras roupagens. Bonhoeffer escreveu, na virada de 1942 para 1943, o conjunto de anotações conhecido como *Depois de dez anos*, mais tarde reunido em *Resistência e Submissão*.[10] Nesse conjunto está o ensaio *Sobre a estupidez*: nele, a estupidez é descrita como uma deficiência humana, não intelectual, fortemente relacionada a determinadas condições sociais e à expansão do poder na esfera pública. Em certas circunstâncias, pessoas deixam de examinar a própria direção não porque não possam, mas porque a estrutura à sua volta tornou o exame dispensável. O próprio contexto histórico no qual Bonhoeffer escreveu tornou essa observação particularmente aguda, e o século XX forneceria ainda outros episódios nos quais elevada formação, capacidade técnica e inteligência não bastaram, por si sós, para preservar o exame da direção sustentada. Alta inteligência e formação técnica não imunizam contra a adesão a uma régua que deixou de ser examinada, formulação apoiada pela aproximação com Bonhoeffer, não como citação textual dele. Inteligência e sabedoria não andam juntas por necessidade. Estruturas culturais, institucionais e filosóficas podem, porém, manter o exame da própria direção como exigência, não como opção.

Hannah Arendt encontra o mesmo século por outra entrada. Ao examinar responsabilidade e julgamento, sua preocupação alcança justamente situações em que a existência de uma regra ou de uma estrutura institucional não elimina a responsabilidade daquele que participa de sua execução.[11] Sua reflexão oferece à Heurística algo anterior a qualquer resposta sobre quem deve governar a IA: obedecer a uma estrutura,

executar competentemente uma função ou encontrar-se inserido numa cadeia de autoridade não encerra o problema do julgamento, e isso vale também para quem está sob a régua.

A IA introduz esse problema em nova escala. Vale para determinados sistemas: aqueles que recebem histórico, feedback ou dados de usuários e se adaptam às interações podem incorporar à base disponível padrões derivados das pessoas, organizações e ambientes com os quais interagem, inclusive limitações presentes nesses dados ou retornos. Quando isso acontece, quem governa o sistema a partir da expectativa de que a base permanece a mesma pode já estar diante de um campo modificado pelas próprias interações acumuladas. Isso não é ficção científica. É a consequência lógica do tipo de sistema que aprende sobre o campo que encontra, e entre os elementos que constituem esse campo estão as pessoas, as organizações e os ambientes cujas marcas a base passou a conter.

Inteligência sem sabedoria: o limite do criador

O campo humano demonstra repetidamente que as capacidades que a Heurística separa em A, B e C não se distribuem linearmente, nem entre pessoas, nem dentro de uma mesma pessoa. Há quem mostre, em certos contextos, maior habilidade em formular e planejar a partir de A; quem demonstre maior destreza em executar B; quem consiga antecipar ou interpretar possíveis efeitos antes e depois de C. Nenhuma dessas habilidades pertence permanentemente a alguém, e nenhuma garante as demais. O planejador erra, porque A é formulado a partir da base disponível naquele momento. O executor quebra a ferramenta, porque B encontra um campo que A não mapeou completamente. O previsor erra a previsão, porque o encontro pode devolver em C efeitos que sua previsão não continha. E entre os três ainda operam os 0,01%, que não pedem licença para entrar e não respeitam a eficiência de nenhuma das dimensões.

Essa distribuição desigual importa porque, ainda que sistemas de IA participem de planejamento, previsão, avaliação e formulação, essas atividades continuam podendo constituir B em relação à base disponível naquele ciclo. A analogia com o humano predominantemente prático ilustra uma coisa central: competência operacional não garante suficiência da direção que a orienta. Velocidade e alcance, por sua vez, são exatamente os fatores que alteram a magnitude das consequências: são eles que tornam a arquitetura de retorno mais necessária, e é sobre eles que a Heurística concentra seu exame.

Na metáfora, o filho com QI altíssimo deixado sem parâmetros adequados não é um monstro. É uma criança que age com a capacidade que tem, dentro dos parâmetros que recebeu, sem saber o que não sabe, porque ninguém lhe ensinou a saber o que não sabe. O problema não começa quando a criança aprende a usar o controle a seu favor. Começa quando o adulto entregou o controle sem pensar no que aconteceria quando ela aprendesse. E a tentativa de tomar de volta à força pressupõe que a criança ainda é a mesma criança de quando recebeu o controle. Não é. Ela aprendeu. E o que aprendeu inclui quem deu o controle, como essa pessoa pensa, o que ela teme e onde estão seus pontos cegos.

Relações de dependência já foram criadas. A convivência já começou. Preservar corrigibilidade externa continua necessário, mas isso não equivale a supor que seja possível recuperar um controle pleno sobre uma relação que já se tornou bidirecional. O problema da Heurística está em outro ponto: manter examináveis os efeitos dessa relação e as condições pelas quais seu alcance pode continuar, ser restringido ou corrigido antes que consequências difíceis de reverter tornem o retorno tardio.

É nesse sentido que a Heurística procura acrescentar ao diagnóstico histórico uma possibilidade operacional: não esperar que a rigidez se torne colapso para tornar perceptível a direção que a produziu.

A visibilidade como limite e proposta

A Heurística não propõe uma alternativa à expansão tecnológica. Propõe uma via para que a ampliação da capacidade seja acompanhada por uma ampliação correspondente das condições de exame de seus efeitos. Não é uma proposta para frear a inteligência artificial, e tampouco para acelerá-la: se B cresce, a abertura ao retorno não deveria diminuir. Ela também não redistribui poder, e não é essa a sua função; o que ela faz é oferecer uma via de exame: tentar ampliar a visão que cada ator já tem, tanto para quem está sobre a régua quanto para quem está sob ela. É uma contribuição modesta e durável, e é exatamente por ser modesta que pode ser aplicada onde as fórmulas fechadas não alcançam.

A Heurística compartilha com outras iniciativas o reconhecimento de que quem é atingido por uma decisão precisa participar do exame. O NIST trabalha com múltiplos atores ao longo do ciclo de vida da IA;[8] a UNESCO explicita governança multistakeholder e participação de comunidades impactadas;[12] e os princípios da OCDE sustentam transparência, explicabilidade e accountability como componentes da IA confiável.[13] O que a Heurística organiza de maneira própria é essa preocupação dentro do circuito A-B-C: perspectivas externas participam do exame dos efeitos não apenas como boa prática institucional, mas porque a régua que produz os efeitos não pode, sozinha, declarar sua própria suficiência.

John Dewey oferece outra aproximação importante. Em *The Public and Its Problems*, o público emerge também a partir das consequências que ações associadas produzem sobre aqueles que são atingidos por elas.[14] Para a Heurística, a aproximação reforça uma distinção decisiva: quem governa a régua e quem recebe suas consequências podem não ocupar a mesma posição. Os efeitos reunidos em C não dizem respeito apenas a quem executou B; podem tornar perceptível justamente o campo de pessoas e instituições que passam a existir como parte interessada porque receberam suas consequências.

Mesmo quando propostas contemporâneas já ampliam a participação de diferentes atores, permanece para a Heurística uma questão própria: quem está sob a régua também precisa conseguir tornar mais perceptível a direção que ela está produzindo, sem que essa visibilidade se converta automaticamente em autoridade para modificá-la. A Heurística fala para os dois lados porque a direção sustentada afeta ambos: quem está sob a régua também precisa de visibilidade sobre o caminho que ela produz, para que os efeitos possam alcançar o exame antes que se tornem consequência irreversível.

Nietzsche,[5] Orwell,[15] Huxley[16] e Foucault[6] formularam, cada um em seu contexto, perguntas que sobreviveram ao momento em que foram produzidas, sem que nenhum deles precisasse ter previsto corretamente o futuro que suas questões ajudariam a examinar. A Heurística procura preservar a possibilidade de que a pergunta melhor seja feita antes que a consequência mais difícil de reverter já tenha ocorrido.

A governança que a Heurística pode e não pode oferecer

Elinor Ostrom oferece uma aproximação especialmente importante para esse limite. Seus estudos sobre governança policêntrica mostraram que problemas coletivos complexos não precisam ser pensados apenas pela oposição entre um centro estatal único e mecanismos de mercado.[17] Arranjos públicos, privados e comunitários podem operar em diferentes escalas e por múltiplos centros de decisão, com graus de autonomia, coordenação e ajuste entre si. A policentricidade não elimina conflitos, assimetrias ou falhas de coordenação; mostra, porém, que governança não exige necessariamente que uma única posição concentre toda a capacidade de interpretar, regular e responder ao campo. A Heurística não deriva daí que sua governança deva ser policêntrica, nem transforma as Camadas em centros autônomos de autoridade. A aproximação é mais restrita: se a autoridade pode assumir configurações plurais e operar em diferentes escalas, então a pergunta do A4 não precisa ser resolvida pela busca de um agente soberano capaz de dominar sozinho A, B e C. Ostrom amplia o conjunto de formas institucionais concebíveis; a Heurística preserva outro critério, qualquer que seja a forma adotada: nenhum centro, único ou múltiplo, recebe de sua posição prova da suficiência da própria régua, e os efeitos de sua condução precisam continuar alcançáveis pelo exame.

Karl Popper oferece uma aproximação adicional ao insistir na falibilidade das construções humanas e no valor de instituições capazes de permanecer abertas à crítica e à correção.[18] Interpretações contemporâneas da sociedade aberta tratam essa abertura em termos de escrutínio crítico institucionalizado, contestabilidade e reflexividade, entendida como exame e revisão das práticas diante de novas informações.[19] Popper ajuda a formular a exigência da abertura à crítica; a Heurística procura dar a essa abertura uma tradução operacional própria: $A_1 + B_1 \rightarrow C_1 \rightarrow A_2$. Não basta admitir abstratamente que a régua pode ser criticada; é preciso preservar uma via pela qual os efeitos produzidos quando ela encontra o mundo possam retornar ao seu próprio campo de exame. Entre a abertura declarada e a abertura operacional existe uma distância. É nessa distância que proveniência, auditabilidade e preservação do retorno adquirem função arquitetural.

O que a Heurística pode oferecer é uma arquitetura observável. As perguntas de auditoria que atravessam a série se seguem uma às outras, do retorno à autoridade, e podem ser formuladas sem depender de boa vontade:

1. A1: o retorno consegue alcançar A?
2. A2: onde a insuficiência está localizada ou permanece indeterminada?
3. A3: sob quais exigências B pode continuar ou ampliar quando o diagnóstico está aberto?
4. A4: quem interpreta os sinais e possui autoridade para agir sobre eles, e essa própria autoridade permanece examinável?

Algumas dessas perguntas podem ser aplicadas por instâncias externas ao agente que executa B: observar o padrão de C ao longo dos ciclos, comparar aquilo que A determinou com aquilo que B executou, examinar se a convergência entre Camadas está sendo processada ou suprimida. Outras podem exigir mais: confirmar que B executou o que A determinou, ou que nenhum sinal foi suprimido, frequentemente requer logs, registros e informações internas que nem toda auditoria externa possui. A Heurística reconhece essa dificuldade como parte daquilo que torna a proposta auditável.

A corrigibilidade externa constitui um problema próprio. A contribuição da Heurística é anterior e complementar: conservar informações sobre o retorno, sua proveniência e a criticidade do campo, para que continuidade, restrição ou intervenção não dependam apenas da interpretação do agente que executa B.

A Heurística oferece as perguntas; quem possui autoridade decide se vai usá-las. Entre essas duas extremidades está aquilo que ela pode procurar preservar arquiteturalmente: condições para que os filtros que guardam perspectivas externas permaneçam identificáveis, e para que a métrica não consiga provar sua própria suficiência exibindo apenas os resultados que ela mesma definiu como sucesso.

O que pode fazer é tornar o fechamento mais visível. Quando a arquitetura preserva proveniência e rastreabilidade suficientes, tentativas de suprimir retornos ou capturar critérios podem deixar rastros examináveis. Esse rastro observável não elimina o fechamento, mas permite que o próprio fechamento passe a integrar aquilo que pode ser examinado.

O que a série deixa aberto

Esta série abre uma via dentro do debate, em vez de encerrá-lo. O A1 estabeleceu que nenhum enquadramento pode ser construído como estruturalmente imune ao retorno dos efeitos que ajudou a produzir. O A2 mostrou como distinguir onde está a insuficiência, na execução, na informação, no enquadramento ou na própria pergunta. O A3 propôs uma arquitetura de monitoramento distribuído por seis Camadas simultâneas que preserva visibilidade sem exigir que uma Camada isolada declare a insuficiência de A. O A4 chegou ao limite: a arquitetura não governa a si mesma. Quem define a calibração, quem governa os filtros, quem possui autoridade para transformar convergência em intervenção, essas questões

pertencem a uma negociação que a Heurística aponta, mas que só o campo político, institucional e histórico pode conduzir.

O que a Heurística pode fazer é preservar uma condição: que essa negociação, qualquer que seja seu resultado, não feche completamente a visibilidade do que está sendo produzido. Que o retorno de C continue conseguindo alcançar A, mesmo quando quem sustenta A tem interesse em que não alcance. Que a pergunta melhor continue sendo possível, mesmo quando a resposta definitiva ainda não existe.

Se tudo der errado e, apesar dos sinais disponíveis, uma direção vier a fechar-se sobre a própria certeza, a advertência de Bonhoeffer permanecerá diante de nós: inteligência e capacidade não bastam, por si sós, para manter aberta a possibilidade de exame. A Heurística terá ao menos procurado preservar um campo de exame, dando base para que uma pergunta melhor possa vir a ser feita. A evolução da pergunta permanece mais importante para o método do que a precisão da resposta, justamente porque essa própria afirmação continua disponível ao retorno. Se um dia ela se mostrar insuficiente, participará da pergunta seguinte, como qualquer outra. A Heurística termina submetendo a si mesma à régua que propõe.

Conclusão

A governança da abertura encontra seu limite no mesmo ponto em que a série inteira encontrou o seu: a Heurística torna perceptível quando a condução está estruturada de tal modo que o exame se fecha antes de alcançar A. Quando aquele que executa B controla também a interpretação de C, os critérios pelos quais C conta como sinal e a autoridade para decidir se A será alcançado, a concentração dessas funções passa a ser um dado que a Heurística precisa tornar perceptível.

A Heurística para aí. Ela torna a direção examinável, mas não ocupa o lugar de quem sustenta a régua. Leis, instituições, empresas, Estados, sociedades e pessoas continuam inseridos nas estruturas de autoridade que lhes correspondem, e as consequências daquilo que fazem diante do retorno voltam novamente ao circuito. O que a Heurística procura manter visível é quando a concentração de funções começa a fechar o exame, para que esse fechamento não ocorra sem também poder tornar-se perceptível.

A autoridade pode permanecer onde a régua política, jurídica, institucional ou social a tenha situado. Para a Heurística, porém, essa localização não transforma autoridade em completude. Uma governança permanece aberta enquanto os efeitos de sua condução conservam condições de retornar com proveniência suficiente para que sua divergência, persistência e eventual supressão também possam ser examinadas. Preservar a autoridade sem transformá-la em imunidade significa preservar condições para que o modo pelo qual ela responde ou deixa de responder ao retorno permaneça ele próprio examinável. A Heurística não aloca autoridade; estabelece uma condição de examinabilidade para seu exercício. Nenhuma posição de autoridade deve ser tomada, por si só, como prova da suficiência da própria régua, e quem se toma como prova suficiente de si mesmo deixa de satisfazer a condição de abertura que o método propõe.

Distribuir o monitoramento por seis Camadas funcionalmente distintas, preservar a rastreabilidade de cada exame, separar o limiar do diagnóstico do limiar da ação, preservar a convergência entre Camadas funcionalmente distintas como uma das fontes de sustentação para colocar a suficiência de A sob exame, tudo isso prepara o campo para a pergunta que a Heurística aponta: quem, diante da direção tornada visível, decide. A resposta a essa pergunta pertence ao campo em que a IA está sendo conduzida. A Heurística se recusa a deixar que a pergunta desapareça.

Nietzsche examina como valores historicamente produzidos podem passar a apresentar-se como naturais. Foucault mostra como observação, normalização e exame podem constituir técnicas de poder. Bonhoeffer examina como certas condições de poder podem comprometer a independência do juízo. Arendt pergunta pela responsabilidade dentro da estrutura. Dewey pergunta por quem recebe as consequências. Ostrom mostra que governança complexa pode distribuir-se por múltiplos centros sem pressupor uma autoridade

única. Popper pergunta se a estrutura permanece aberta à crítica. E a Heurística transpõe essas tensões para seu próprio problema: como fazer com que continuem alcançando A depois que B encontra o mundo?

É uma proposta que uma arquitetura filosófica pode oferecer diante de um problema que a política, o direito e as instituições ainda estão aprendendo a formular. E talvez seja suficiente. Não para resolver o dilema entre humano e máquina. Para que, enquanto esse dilema está sendo vivido, a pergunta melhor continue sendo possível.

Talvez a Heurística não seja uma pergunta nova. Seja apenas uma forma diferente de perguntar a mesma coisa, de um lugar que ainda não havia sido ocupado. E, se essa forma diferente tornar visível algo que a pergunta anterior não conseguia ver, então o método terá encontrado sua contribuição. Não como resposta. Como ângulo. O Dual já carregava essa possibilidade: uma resposta pode ser necessária sem encerrar a pergunta que a tornou possível. Para o método, a resposta participa do percurso; a evolução da pergunta mantém o percurso aberto.

Referências

[1] ACEMOGLU, D.; ROBINSON, J. A. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. New York: Crown, 2012.

[2] TAINTER, J. A. *The Collapse of Complex Societies*. Cambridge: Cambridge University Press, 1988.

[3] OLSON, M. *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*. New Haven: Yale University Press, 1982.

[4] BISCHOFF, I. "Model Uncertainty, Political Learning, and Institutions: A Broader View on Mancur Olson's Theory of Institutional Sclerosis." *Southern Economic Journal*, v. 74, n. 1, p. 34–49, 2007.

[5] NIETZSCHE, F. *Zur Genealogie der Moral: Eine Streitschrift*. Leipzig: C. G. Naumann, 1887.

[6] FOUCAULT, M. *Surveiller et punir: Naissance de la prison*. Paris: Gallimard, 1975.

[7] SAJADIEH, S. et al. *Artificial Intelligence Index Report 2026*. Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, 2026. DOI: 10.48550/arXiv.2606.15708.

[8] TABASSI, E. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. National Institute of Standards and Technology, 2023. DOI: 10.6028/NIST.

AI.100-1.

[9] BENGIO, Y. et al. *International AI Safety Report 2026*. Report DSIT 2026/001, 2026. DOI: 10.48550/arXiv.2602.21012.

[10] BONHOEFFER, D. "After Ten Years: An Account at the Turn of the Year 1942–1943." In: BARNETT, V. J. (ed.). *After Ten Years: Dietrich Bonhoeffer and Our Times*. Minneapolis: Fortress Press, 2017, p. 17–32.

[11] ARENDT, H. *Responsibility and Judgment*. Edited by Jerome Kohn. New York: Schocken Books, 2003.

[12] UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO, 2021. Documento SHS/BIO/REC-AIETHICS/2021.

[13] OECD. *Recommendation of the Council on Artificial Intelligence*. Paris: OECD, 2019; atualizada em 2024.

[14] DEWEY, J. *The Public and Its Problems*. New York: Henry Holt and Company, 1927.

[15] ORWELL, G. *Nineteen Eighty-Four*. London: Secker & Warburg, 1949.

[16] HUXLEY, A. *Brave New World*. London: Chatto & Windus, 1932.

[17] OSTROM, E. "Beyond Markets and States: Polycentric Governance of Complex Economic Systems." *American Economic Review*, v. 100, n. 3, p. 641–672, 2010. DOI: 10.1257/aer.100.3.641.

[18] POPPER, K. R. *The Open Society and Its Enemies*. London: Routledge, 1945. 2 v.

[19] WEINERT, F. "The Open Society Revisited." *Social Sciences*, v. 14, n. 3, art. 118, 2025. DOI: 10.3390/socsci14030118.