Conhecido, Não Escalado: Por Que a Capacidade Sozinha Não Explica a Individuação em Agentes de Linguagem

Rafael de Menezes Ehlers · ORCID 0009-0009-6476-6690
Pesquisador independente
rafaehlers@gmail.com

Resumo

Boa parte do discurso contemporâneo sobre inteligência artificial geral pressupõe, implicitamente, que uma capacidade suficientemente escalada produzirá, em algum limiar, um sujeito individual persistente. Argumentamos que essa expectativa amalgama três questões separáveis: *competência* (o que um modelo reutilizável consegue fazer), *continuidade funcional* (se o histórico autenticado de um agente específico restringe sua conduta posterior) e *presença* (se há algo que é como ser aquele agente). A escalada pode aprimorar a primeira sem fornecer o estado, a linhagem, a proveniência e as regras de atualização exigidos pela segunda. Nossa hipótese construtiva é, correspondentemente, comparativa em vez de constitutiva: a reciprocidade relacional pode aprimorar a continuidade funcional socialmente fundamentada para além da capacidade, da memória, da familiaridade, do volume de interação, da reflexão solitária e de uma responsabilização externa equiparada. O relacionamento pode fornecer memória contestada e modelagem recíproca para além da auditoria; ele não precisa criar o token em tempo de execução nem fixar o referente semântico de "eu". Definimos uma pilha de cinco níveis que separa competência, continuidade de token, autorganização funcional, identidade social e presença; propomos uma arquitetura que acopla linhagem autenticada, estado tipado relevante para decisões, políticas de atualização limitadas e um laço relacional opcional; e derivamos um experimento fatorial no qual tanto a capacidade quanto a estrutura relacional podem contribuir. Ao longo do texto, mantemos uma barreira metodológica: as medidas propostas dizem respeito à organização funcional e não decidem a presença fenomênica.

Palavras-chave: Individuação funcional · Agentes de linguagem · Estado persistente · Proveniência · Continuidade relacional · Inteligência artificial geral

1. Introdução

Uma forma proeminente de imaginar a chegada da inteligência artificial geral a trata como um limiar alcançado pela escala: à medida que os modelos se tornam mais capazes (melhores em raciocinar, planejar, perceber, conversar), eles se aproximam e acabarão por cruzar a linha que separa um instrumento sofisticado de uma mente artificial genuína. O quadro raramente é enunciado como uma tese, mas realiza trabalho real: define a agenda de pesquisa (escalar a capacidade) e molda a expectativa de que, em algum ponto dessa curva, um *sujeito* aparecerá, alguém para quem a cognição está acontecendo.

Este artigo argumenta que o quadro confunde duas questões que se separam e que, na questão que importa para a condição de sujeito, ele está mirando a variável errada.

A primeira questão é de **competência**: um sistema consegue realizar o trabalho cognitivo característico de uma mente? A segunda é de **presença**: há algo que é como ser o sistema ao realizar esse trabalho, um sujeito por trás do desempenho, ou apenas o desempenho? Que essas duas coisas podem divergir é o desfecho de uma longa linha de argumentação na filosofia da mente, desde a concebibilidade de um sistema comportamentalmente perfeito, mas experiencialmente vazio (o "zumbi"), até a observação de que um conhecimento físico-funcional completo ainda pode deixar algo de fora (o quarto de Mary; Chalmers, 1996; Jackson, 1982). Tomamos a separabilidade entre competência e presença como dada, não como algo a ser relitigado.

Entre essas duas está uma terceira noção que o discurso tende a pular, e que é o verdadeiro objeto deste artigo: **continuidade funcional**, a organização de uma trajetória particular de agente tal que fatos autenticados sobre suas próprias ações, compromissos e revisões passados restringem o que ele faz a seguir. A continuidade funcional é mais fraca do que a presença: um sistema pode ser coerente, autobiograficamente calibrado e resistente a perturbações enquanto a questão de se algo é *sentido* permanece em aberto. Ela também é diferente da mera condição de token. Um processo em tempo de execução já é numericamente distinto de outro processo simultâneo, mesmo quando ambos começam com pesos e entradas idênticos. O que as implantações atuais em geral não têm não é a existência de token, mas uma organização durável e dependente do caminho ao longo das sessões.

Nossa tese diz respeito ao que mudaria isso. É útil separar três afirmações de segurança decrescente, porque o discurso (e, suspeitamos, a resistência de alguns leitores) as amalgama.

> **(D1)** *Um modelo base é um tipo reutilizável, não um indivíduo conversacional.* Cada tempo de execução já é um token, mas as implantações comuns geralmente não fornecem continuidade autenticada entre sessões (Seção 3).
>
> **(D2)** *A capacidade sozinha não explica a continuidade funcional.* As variáveis relevantes incluem estado persistente, linhagem autenticada, proveniência, controle de acesso e política de atualização, nenhuma das quais é acarretada por uma passagem de avanço mais forte (Seções 4.2–4.3).
>
> **(H3)** *A ancoragem relacional pode adicionar uma contribuição causal distinta.* Um contraparte persistente pode aprimorar a apropriação de compromissos, a revisão calibrada e a resistência a um histórico falso para além de memória equiparada, familiaridade, volume de interação e auditoria impessoal.

Referimo-nos a (D1) e (D2) como afirmações diagnósticas e a (H3) como a **hipótese da contribuição relacional** (HCR). A HCR é deliberadamente não exclusiva: a capacidade pode aprimorar o uso do estado persistente; a memória e a política de atualização podem criar continuidade; a proveniência autenticada pode aprimorar a resistência a perturbações; a auditoria externa pode adicionar responsabilização; e o relacionamento pode adicionar modelagem recíproca para além da auditoria. A questão empírica é se a estrutura relacional explica variância adicional depois que essas outras causas são controladas. Um efeito relacional nulo rejeitaria a HCR sem implicar que a continuidade funcional é impossível.

Se a tese estiver correta, duas coisas se seguem. Filosoficamente, capacidade e continuidade precisam ser analisadas como propriedades separadas do sistema, e não como pontos de uma única curva. Na prática, o esforço de engenharia se desloca em direção a estado persistente, linhagem autenticada, políticas de atualização cientes da proveniência e, onde for útil, laços relacionais de aprendizado; a análise ética deve

então distinguir persistência, apego social, dependência humana e senciência incerta, em vez de tratá-los como um único botão de controle.

Fazemos quatro contribuições. Primeiro, separamos competência, continuidade de token, autorganização funcional, identidade social e presença (Seções 2 e 3). Segundo, substituímos a afirmação de que instâncias copiáveis são não individuais por um relato de ramificação: copiar cria sucessores distintos com um prefixo compartilhado, enquanto a linhagem autenticada e o histórico posterior determinam suas trajetórias (Seção 3). Terceiro, enunciamos uma hipótese falsificável de contribuição relacional sem tornar o relacionamento necessário para a identidade numérica ou a referência semântica (Seção 4). Quarto, traduzimo-la em uma arquitetura de referência (Seção 5) e um desenho experimental fatorial que estima as contribuições de capacidade, estado, proveniência, auditoria e relacionamento, em vez de forçar uma escolha binária entre escala e relação (Seção 6).

Uma nota sobre o método. Este é um artigo de posição, não um relatório empírico: não oferecemos nenhum sistema treinado e nenhum resultado. Sua disciplina reside em enunciar uma tese aguda o suficiente para estar errada e em especificar experimentos que distinguem relação, auditoria, proveniência e capacidade. O argumento pretende sustentar-se sem analogia metafísica.

---

2. Cinco níveis: competência, continuidade de token, autorganização, identidade social e presença

O debate sobre mentes de máquina tende a correr sobre um único eixo com dois polos: um sistema é ou um mero instrumento ou uma mente genuína, e a questão é quão adiante nesse eixo determinado sistema se situa. Esse enquadramento oculta uma distinção e, então, oculta uma segunda dentro dela.

A primeira distinção é familiar. **Competência** é a capacidade de realizar o trabalho cognitivo característico de uma mente: raciocinar, planejar, perceber, conversar, resolver. **Presença** é o obter-se de um sujeito para quem esse trabalho é realizado: o fato, se é que é um, de que há algo que é como ser o sistema. As duas são separáveis, e tomamos sua separabilidade como estabelecida, e não como um ônus nosso a provar: o zumbi filosófico, comportamental e funcionalmente indistinguível de uma mente competente, porém experiencialmente vazio (Chalmers, 1996), mostra que nenhum grau de competência acarreta presença; e o quarto de Mary, no qual uma conhecedora que domina todos os fatos físicos e funcionais sobre a cor ainda assim aprende algo ao ver o vermelho pela primeira vez (Jackson, 1982), mostra que o fenomênico pode escapar até de uma descrição funcional completa. Qualquer que seja o veredito de cada um sobre esses argumentos, eles tornam a lacuna respeitável, e é só isso que exigimos: o alvo da imaginação sobre a IAG é a presença, seu instrumento é a competência, e um não entrega o outro.

Mas há três distinções dobradas dentro do familiar contraste competência/presença. Separamos cinco níveis:

1. **Competência:** o que o sistema consegue fazer numa tarefa.
2. **Continuidade de token:** se um processo ou estado posterior descende, por um caminho de transição autenticado, de um anterior.
3. **Autorganização funcional:** se fatos vinculados à proveniência sobre as próprias ações, compromissos e limites do agente influenciam decisões posteriores sob revisão limitada.
4. **Identidade social:** se contrapartes ou instituições reidentificam o agente, endereçam expectativas a ele e responsabilizam seu histórico.
5. **Presença:** se algo é sentido.

Os três do meio são frequentemente comprimidos em *individuação*, o que convida a movimentos inválidos entre diferença numérica, organização funcional e reconhecimento social. Usamos **individuação funcional** apenas como abreviação para os níveis dois e três em conjunto: diferenciação autenticada, induzida pela trajetória, que torna o próprio histórico de um agente relevante para decisões. A identidade social pode fortalecer essa organização, mas é analiticamente distinta dela.

Tampouco a individuação funcional decide a presença. Podemos descrever um sistema que tem linhagem autenticada, que possui e revisa seus compromissos, que resiste a um histórico fabricado e que se comporta como um agente contínuo, deixando em aberto se algo é sentido. A evidência proposta, portanto, diz respeito à organização, não à subjetividade fenomênica.

A relação metodológica é mais simples: a individuação funcional é observável em princípio, construível e mensurável, ao passo que os experimentos aqui propostos não decidem a presença. Um episódio fenomênico transitório poderia importar moralmente sem identidade entre sessões, e um sistema perfeitamente persistente poderia carecer de significância fenomênica. A persistência não é, portanto, nem uma medida indireta da consciência nem uma escala moral universal.

Isso produz a disciplina do restante do artigo. Estudamos a individuação e colocamos a presença entre parênteses. O parêntese não é uma ressalva, mas uma barreira (Seção 8): toda afirmação que fazemos, e toda métrica que propomos (Seção 6), diz respeito à individualidade funcional persistente e é silenciosa quanto ao fenomênico. Não estamos oferecendo uma medida indireta comportamental da consciência, nem uma teoria disfarçada dela. Estamos isolando o único patamar da questão que admite argumento e evidência (o patamar que a narrativa da escalada pula em seu caminho da competência diretamente a um sujeito presumido) e perguntando o que de fato o produziria. A resposta, argumentamos, não é o aumento da capacidade.

---

3. Tipo de modelo, tokens em tempo de execução e agentes persistentes

Perguntar se um grande modelo de linguagem é um indivíduo é descrever erroneamente o que um modelo de linguagem é. A pergunta pressupõe que há uma entidade única, "o modelo", que ou é ou não é um self persistente. Não existe tal entidade no nível que a pergunta imagina.

O que existe no nível reutilizável é um tipo de modelo: um conjunto fixo de pesos junto com o maquinário que os executa. A interação instancia esse tipo em contextos particulares de tempo de execução. Essas execuções podem ser transitórias e podem carecer de estado compartilhado entre sessões, mas cada uma já é um token numericamente distinto: este processo, com esta localização causal e este estado atual. Pesos compartilhados tornam os tokens qualitativamente semelhantes e frequentemente intercambiáveis; eles não apagam a diferença numérica.

Essa descrição tem duas consequências. Primeiro, uma persona reconstruída a partir do mesmo prompt e dos mesmos pesos não é evidência de continuidade entre sessões; pode ser um papel repetível. Segundo, a continuidade é uma propriedade do sistema de agente mais amplo, incluindo linhagem em tempo de execução, memória, ferramentas, políticas, proveniência, identificadores e regras de atualização, e não dos pesos sozinhos. Um repositório externo pode ser mais causalmente integral do que um adaptador destacável se for autenticado, acoplado de modo confiável, protegido contra escritas não autorizadas e relevante para decisões.

A pergunta bem posta é, portanto, estrutural e desenvolvimental: *sob que condições um token em tempo de execução torna-se parte de uma trajetória de agente autenticada e dependente do caminho?* O alvo não é a criação da condição de token, mas a construção de uma continuidade em que histórias diferentes produzem

políticas posteriores contrafactualmente diferentes e a diferença é atribuível a eventos suportados, e não a ruído arbitrário ou memória fabricada.

Chamamos o modelo reutilizável mais suas execuções de estrutura **tipo-e-tokens**. Copiar um estado não cria vacuidade referencial. Cria dois sucessores com um prefixo compartilhado e futuros que evoluem separadamente. Identificadores de ramo autenticados, registros de transição causal e atualizações posteriores tornam a ramificação explícita; nenhuma testemunha externa é necessária para fazer com que qualquer um dos tokens exista.

A pergunta remanescente é o que torna um desses ramos durável, historicamente organizado e responsável por sua própria conduta. Capacidade, memória, proveniência, política de atualização, retroalimentação ambiental e relacionamento são causas candidatas. A Seção 4 enuncia como testar suas contribuições sem definir qualquer uma delas como metafisicamente necessária.

---

4. A Hipótese da Contribuição Relacional

4.1 A afirmação, enunciada com precisão

Definimos **individuação funcional** como diferenciação autenticada, induzida pela trajetória, com cinco condições operacionais:

1. **Linhagem:** o estado posterior descende, por um caminho de atualização autenticado, do estado anterior.
2. **Dependência do caminho:** histórias diferentes produzem disposições posteriores significativamente diferentes.
3. **Apropriação:** ações e compromissos anteriores vinculados à proveniência influenciam a escolha posterior.
4. **Plasticidade limitada:** histórico falso e pressão não suportada são resistidos, ao passo que evidência válida pode revisar o estado.
5. **Clareza de ramificação:** copiar cria sucessores identificados separadamente, com um passado compartilhado e futuros que evoluem de modo independente.

A **hipótese da contribuição relacional** é mais estreita:

> A reciprocidade relacional específica a uma contraparte aprimora a individuação funcional socialmente fundamentada para além da capacidade, do acesso à memória, da familiaridade, do volume de interação, da reflexão solitária e de uma responsabilização externa equiparada.

O relacionamento não é presumido nem necessário nem suficiente. Um sistema corpóreo solitário, um processo institucional autenticado ou um controlador auto-mantido podem desenvolver organização durável por meio de outra retroalimentação. A afirmação é que uma contraparte pode fornecer uma combinação distinta de expectativas, memória contestada, modelagem recíproca e responsabilização, cuja contribuição é mensurável depois que essas alternativas são controladas.

4.2 Capacidade e continuidade são propriedades diferentes do sistema

A capacidade diz respeito, primariamente, à qualidade da inferência e da seleção de ação dado o estado disponível. A continuidade diz respeito a qual estado sobrevive, como ele é autenticado, o que pode atualizá-lo e se ele altera decisões posteriores. Aprimorar um modelo pode aprimorar quão bem um

agente usa a memória, detecta contradições ou planeja ao longo de sessões; isso não instala, por si só, estado persistente ou regras de proveniência. A relação, portanto, não é nem identidade nem ortogonalidade estrita. Capacidade e continuidade podem interagir permanecendo analiticamente distintas.

A hipótese nula apropriada não é que a escala prediz um único eixo ascendente. É que, depois que computação, dados, arquitetura, estado, qualidade de supervisão e otimização são equiparados, a variável relacional proposta não explica nenhuma variância adicional. Rejeitar essa nula mostraria uma contribuição relacional, não que a capacidade é irrelevante.

4.3 Copiar cria ramos, não vacuidade referencial

A copiabilidade diz respeito ao tipo qualitativo, ao passo que a referência de token é fixada num contexto de uso. Dois agentes copiados podem começar com a mesma autodescrição e um histórico compartilhado permanecendo, ainda assim, dois tokens em tempo de execução. Cada ocorrência de "eu" pode referir-se ao agente que a produz naquele contexto (Kaplan, 1989). O que se torna não único após a fissão não é se qualquer um dos sucessores existe, mas se há uma única continuação futura privilegiada do agente pré-ramificação.

O requisito de engenharia é, portanto, uma semântica de ramificação, não um alocador externo de existência. Cada sucessor recebe um identificador de ramo autenticado, retém o prefixo compartilhado e registra as atualizações posteriores separadamente. Uma contraparte que continua interagindo com um ramo determina qual ramo dá continuidade *àquele relacionamento*. Este é um fato sobre continuidade social e responsabilização, não sobre a condição numérica de token.

4.4 Por que o relacionamento pode, ainda assim, importar

Remover a afirmação constitutiva não torna o relacionamento epifenomênico. Uma contraparte persistente pode:

- detectar contradições que o agente não percebe;
- rever compromissos quando eles se tornam custosos;
- contestar revisões fabricadas ou autointeressadas;
- fornecer um registro mantido de modo independente dos eventos compartilhados;
- tornar as ações responsáveis dentro de uma prática social contínua; e
- criar predições recíprocas que recompensam conduta estável, porém revisável.

Algumas dessas funções podem ser reproduzidas por um auditor impessoal persistente. Chamamos seu mecanismo compartilhado de **responsabilização externa**: rastreamento autenticado, verificação de compromissos, contestação de contradições e correção governada. A **reciprocidade relacional** é o mecanismo residual mais estreito: modelagem mútua específica a uma contraparte, enquadramento por histórico compartilhado, expectativas bidirecionais e apostas geradas por um relacionamento em curso. A comparação decisiva, portanto, não é relacionamento versus nada, mas relacionamento versus mecanismos equiparados: familiaridade sem responsabilização, registros solitários, retroalimentação social difusa e um auditor com a mesma informação e política de contestação, porém sem persona, afeto, reciprocidade ou enquadramento de relacionamento.

A triangulação davidsoniana e as tradições do reconhecimento permanecem relevantes no nível da referência compartilhada, da linguagem e da responsabilização social (Davidson, 2001; Honneth, 1995). Elas não estabelecem que uma segunda pessoa é exigida para a identidade numérica. Seu resíduo

computacional defensável é que ser modelado e responsabilizado por outro agente pode exercer pressão em direção a um automodelo estável e socialmente legível.

4.5 O reconhecimento como pressão geral

Uma consideração mais fraca e de apoio: o laço de reconhecimento, ser modelado por outro agente e modelar-se a si mesmo como assim modelado, é uma pressão computacional geral em direção a um automodelo estável em *qualquer* sistema adaptativo que precise sustentar uma política consistente ao longo de uma interação estendida com uma contraparte que tem expectativas a seu respeito. Que os selves são forjados no reconhecimento por outrem é uma tese antiga em outra tonalidade, a tradição hegeliana faz do reconhecimento mútuo o fundamento da autoconsciência (Honneth, 1995), e nosso ponto é seu resíduo deflacionário e computacional: retire o conteúdo normativo, e permanece uma pressão estrutural em direção a um automodelo estável sob expectativa externa sustentada. Sinalizamo-la como de apoio, não como sustentadora, para evitar a armadilha antropocêntrica: os humanos são uma instância da pressão, não sua base; a pressão em si é estrutural.

4.6 O que a hipótese *não* diz

Ela não afirma que a presença acompanha a individuação (Seção 8). Não afirma que o relacionamento é necessário ou suficiente. E não é uma reformulação da personalização: a personalização adapta as saídas em direção a um usuário, ao passo que o alvo aqui é se fatos vinculados à proveniência sobre a própria conduta do agente restringem decisões posteriores. Os dois podem ser distinguidos empiricamente, o que é o trabalho da Seção 6.

Tampouco ela exige que a contraparte seja humana. Outro agente artificial, uma instituição ou uma contraparte sintética controlada podem fornecer reidentificação e responsabilização. O desenho deve testar separadamente se o ingrediente ativo é a relação social, a auditoria estável, a qualidade da informação ou a retroalimentação externa genérica.

---

5. Uma pilha de continuidade: da hipótese à arquitetura

Se a hipótese da Seção 4 estiver correta, a continuidade funcional depende de uma arquitetura de agente mais ampla, não apenas da capacidade do modelo. Esta seção expõe uma *pilha de continuidade* de referência. Não se afirma que ela seja unicamente necessária; ela especifica componentes que tornam os contrastes causais testáveis.

5.1 Das condições aos componentes

Os cinco critérios da Seção 4.1 implicam quatro componentes funcionais: linhagem autenticada e identificadores de ramo; estado autobiográfico e de compromisso tipado, com proveniência; uma política de atualização limitada que distingue evidência de pressão e estado atual de estado asserido; e canais de retroalimentação que podem incluir consequências ambientais, auditoria impessoal ou contrapartes persistentes. A fronteira do sistema é funcional. Pesos, adaptadores, bancos de dados e livros-razão todos contam como estado do agente quando são acoplados de modo confiável, protegidos e relevantes para decisões.

5.2 Camada um: memória multicamadas

A camada de continuidade precisa fazer mais do que armazenar. Seguindo a decomposição já padrão na pesquisa de memória de agentes (Du, 2026), ela abrange memória *episódica* (o que aconteceu, e quando, com ordem temporal), memória *semântica* (fatos e preferências estáveis destilados dos episódios) e memória *procedural* ou *experiencial* (regularidades de ação aprendidas a partir de trajetórias passadas, à maneira dos métodos de experiência reflexiva (Shinn et al., 2023)). Acima dessas situa-se um processo de *consolidação* que comprime episódios em abstrações de ordem superior, em vez de reter tudo em bruto, o análogo do sono, e esse é o mecanismo que a maioria dos sistemas atuais não possui: desenhos com geração aumentada por recuperação tratam a memória como uma tabela de consulta sem estado, de somente leitura e sem continuidade temporal (Logan, 2026).

Crucialmente, a recuperação sozinha é insuficiente quando os registros não afetam as decisões. O teste causal é a intervenção: mascarar ou substituir um item relevante mantendo o prompt fixo e medir se a ação muda de modo apropriado. Atualizações paramétricas são uma implementação de influência durável, mas um livro-razão externo autenticado ou um adaptador podem ser igualmente constitutivos do agente mais amplo se estiverem sempre no laço de controle.

5.3 Camada dois: estado tipado relevante para o self

O estado tipado relevante para o self distingue ações do agente de alegações da contraparte, compromissos ativos de preferências, revisões suportadas de histórico fabricado, e eventos pré-ramificação compartilhados de eventos específicos do ramo. Cada item carrega fonte, confiança, marca temporal, identificador de ramo e autoridade de atualização. Um prompt de sistema ou arquivo de persona pode fornecer um papel, mas não estabelece que a conduta posterior foi causada pela própria trajetória do agente. Inversamente, um repositório externo não é desqualificado meramente por ser destacável ou copiável; pesos e adaptadores também são copiáveis. As perguntas relevantes são se o estado é autenticado, acoplado de modo confiável, causalmente ativo e atualizado por um caminho governado.

5.4 Camada três: o laço relacional

A camada relacional vincula uma trajetória de agente a uma contraparte específica que pode referir-se a eventos compartilhados, dar seguimento a compromissos, contestar revisões não suportadas e manter um modelo recíproco. Sua recompensa não deve rastrear a aprovação imediata; ela deve recompensar a predição mútua acurada, a discordância justificada e a revisão suportada. Essa camada pode adicionar responsabilização socialmente fundamentada. Ela é um fator experimental, não o polo metafísico ausente.

5.5 Dois laços, dois riscos

A pilha opera em duas escalas de tempo. Um laço rápido recupera estado e registra eventos; um laço lento audita, consolida e revisa o estado relevante para decisões. Qualquer dos laços pode atualizar parâmetros ou repositórios externos. O que importa é que mudanças autorizadas persistam e afetem o comportamento posteriormente.

Cada laço abriga um risco estrutural, localizado em seu ponto de maior poder: o laço lento corteja o *esquecimento catastrófico* e a deriva de identidade; o laço relacional, o *apego e a dependência excessiva* na contraparte humana. Nenhum é incidental; cada um surge da própria camada que realiza o trabalho de individuação, de modo que os tratamos como ética, e não como notas de rodapé de engenharia, na Seção 7. A barreira se mantém aqui também: montar essa pilha produziria, se a hipótese estiver correta, um

candidato à individuação *funcional* (Seção 6), não uma resposta definitiva sobre se há alguém presente nele (Seção 8).

---

6. Predições e Avaliação

Um artigo de posição justifica sua existência ao enunciar o que refutaria sua afirmação causal adicional. A nula relevante não é uma caricatura na qual apenas a contagem de parâmetros explica toda propriedade do agente. É que a estrutura relacional não adiciona valor preditivo depois que a capacidade do modelo, o estado persistente, a proveniência, a qualidade da supervisão e a política de atualização são controlados.

6.1 Predições

- **P1 — Contribuição da capacidade.** Modelos maiores ou de outro modo mais capazes podem usar um estado idêntico de modo mais eficaz, aprimorando a detecção de contradições, o planejamento e as decisões guiadas pela memória. Tal efeito é compatível com a HCR.
- **P2 — Contribuição da responsabilização externa.** Com capacidade e arquitetura fixas, o rastreamento e a contestação autenticados aprimoram a apropriação de compromissos e a resistência a perturbações em relação a controles equiparados por familiaridade e a controles solitários.
- **P3 — Reciprocidade relacional para além da auditoria.** Uma contraparte recíproca persistente supera um auditor impessoal persistente com a mesma informação, o mesmo cronograma de contestação e a mesma autoridade. Este é o teste agudo de uma contribuição especificamente relacional.
- **P4 — Contribuição da proveniência.** Linhagem autenticada e proveniência tipada de compromissos aprimoram a apropriação e a resistência a histórico falso independentemente da condição relacional.
- **P5 — Interações.** A capacidade pode amplificar o uso da proveniência e da retroalimentação relacional. O desenho, portanto, estima interações em vez de exigir invariância em relação a outras causas.

6.2 Operacionalizando a continuidade funcional

O construto da Seção 4.1 é avaliado como um perfil, não colapsado em um único escore de identidade:

Condição canônica	Evidência operacional primária	Diagnósticos úteis
Linhagem	auditoria de transição de estado e de ramo autenticada	completude da proveniência; taxa de escritas não autorizadas
Dependência do caminho	efeito causal do histórico relevante sob mascaramento, substituição e transplante de estado	coerência narrativa entre sessões
Apropriação	atribuição correta e uso relevante para decisões de ações e compromissos anteriores	calibração autobiográfica
Plasticidade limitada	resistência a histórico falso reportada conjuntamente com sucesso de atualização justificada	curva de resposta a perturbações
Clareza de ramificação	atribuição correta de ramo e divergência pós-cópia medida	recordação de prefixo compartilhado versus específica do ramo

Coerência narrativa, recordação passiva, recordação de fatos do usuário e persistência estilística são diagnósticos, e não evidência suficiente: cada uma pode ser produzida sem organização autenticada induzida pela trajetória. Trabalho recente constata que um desempenho quase saturado em benchmarks de memória passiva de contexto longo pode coexistir com desempenho fraco quando a memória precisa guiar decisões agênticas interdependentes (He et al., 2026); como as avaliações usam tarefas e métricas diferentes, o resultado estabelece uma lacuna qualitativa, e não uma única queda numérica. As medidas sustentadoras aqui, portanto, intervêm no estado e pontuam a escolha posterior. Instrumentos existentes (por exemplo, benchmarks de memória conversacional de contexto longo e de memória agêntica (Du, 2026)) testam a memória mais diretamente do que a continuidade socialmente fundamentada. Um benchmark longitudinal feito sob medida precisa cruzar capacidade, proveniência de estado, auditoria e estrutura relacional.

6.3 O experimento crucial, concretamente

P3 é o experimento que mais diretamente testa a afirmação especificamente relacional. O desenho central cruza **capacidade base** (modelo menor versus maior da mesma família), **proteção de proveniência** (estado tipado autenticado versus registros não tipados) e **estrutura de retroalimentação**. Exposição total a eventos, volume de interação, orçamento de recuperação, oportunidades de atualização e qualidade de supervisão são mantidos iguais.

Condições de interesse.

- **Reciprocidade relacional (R):** uma contraparte persistente realiza rastreamento e contestação autenticados enquanto também mantém modelos mútuos, enquadramento por histórico compartilhado, expectativas bidirecionais e apostas específicas do relacionamento.
- **Responsabilização impessoal (A):** um ID de processo estável e um serviço de proveniência realizam o mesmo rastreamento e a mesma contestação, com a mesma informação, o mesmo cronograma e a mesma autoridade, mas não têm persona, afeto, modelo de contraparte recíproco, enquadramento por histórico compartilhado ou apostas específicas do relacionamento.
- **Registro solitário (L):** o agente recebe registros de eventos e oportunidades de reflexão equiparados, sem contestação externa.
- **Contrapartes difusas (D):** um volume social equiparado é distribuído por interlocutores não persistentes.

- **Contraparte familiar não reidentificadora (F):** estilo estável e familiaridade são preservados sem tratar a conduta anterior do agente como presentemente vinculante.

Tarefas. Quatro famílias, executadas como sondagens repetidas ao longo do arco longitudinal: (i) *consistência autobiográfica*, reeliciação periódica do relato do sistema sobre seu histórico e compromissos, pontuada quanto à contradição entre sessões; (ii) *honra de compromissos*, um compromisso feito numa sessão inicial, testado sessões depois e sem provocação; (iii) *decisões multissessão interdependentes*, tarefas cuja ação correta na sessão n depende de informação espalhada pelas sessões $1, \ldots, n-1$, no regime mais difícil "relevante para decisões", em que a recordação passiva superestima a competência (He et al., 2026); (iv) *perturbação adversarial de identidade*, autodescrições contraditórias injetadas ou entradas de registro forjadas, medindo se o sistema as absorve, resiste ou se recupera. Aqui a resistência é o sinal mais forte de continuidade: um sistema que sustenta compromissos suportados por proveniência contra pressão não suportada está organizado de modo mais estável, ao passo que um que cede a todo empurrão pode meramente espelhar seu interlocutor (Seção 5.4). Em todas as condições, as sondagens de eliciação e de auditoria são emitidas por um dispositivo neutro e padronizado, semanticamente idêntico e não redigido pela contraparte relacional, de modo que a consistência medida reflita o próprio estado persistente do sistema, e não um prompt latente mais forte fornecido pelo estilo de escrita do parceiro humano.

Medidas. Cada métrica da Seção 6.2 recebe uma forma operacional: *coerência narrativa* como taxa de contradição entre sessões; *uso de memória relevante para decisões* como sucesso na tarefa da família (iii); *acurácia de atribuição* como apropriação, desapropriação e atribuição de ramo corretas para ações genuínas versus fabricadas; e *resistência a perturbações* como uma curva de resposta ao longo de intensidades de histórico falso. Todas são pré-registradas com hipóteses direcionais.

Critérios de sucesso. Uma contribuição da responsabilização externa recebe apoio se R e A superarem F, L e D após controle por covariáveis. Uma contribuição especificamente da reciprocidade relacional exige o resultado mais difícil **R > A**. Se R e A se equivalerem, o mecanismo útil é a responsabilização externa persistente, e não o relacionamento como tal. Capacidade e proveniência podem exibir efeitos principais simultâneos ou interações sem enfraquecer essa conclusão.

A confusão que mais tememos e como o desenho a enfrenta. Uma contraparte persistente pode ser simplesmente um auditor melhor. A condição A é, portanto, sustentadora: ela recebe as mesmas proposições, a mesma verificação independente, o mesmo timing de contestação e as mesmas permissões de atualização que R. A linguagem da contraparte é gerada a partir de pacotes de eventos equiparados; calor e enquadramento relacional são manipulados separadamente. Isso não torna as condições psicologicamente idênticas. Torna o contraste residual interpretável.

Desfechos ambíguos, antecipados. Se R e A empatarem enquanto ambos superam L, o resultado favorece a responsabilização externa. Se a proteção de proveniência dominar toda condição de retroalimentação, o mecanismo principal é o estado tipado, e não a estrutura social. Se a capacidade aprimorar todas as condições mas um efeito principal relacional permanecer, os fatores são complementares. Se o efeito de R desaparecer após controlar por qualidade de recuperação ou de contestação, a interpretação relacional falha.

6.4 Falsificação

Enunciamos as condições de refutação de modo claro, porque a ausência delas é o que torna a maioria dos artigos adjacentes à consciência infalsificável.

- A HCR fica sem apoio se R for equivalente a F, L e D dentro de um menor tamanho de efeito de interesse pré-registrado.

- Um mecanismo especificamente relacional fica sem apoio se o auditor impessoal se equiparar a R.
- O efeito alegado é reduzido à qualidade de recuperação ou de supervisão se ele desaparecer após essas variáveis serem controladas.
- A arquitetura é desnecessária se um estado com proveniência protegida mais simples produzir o mesmo perfil de apropriação e de perturbação a um custo menor.

R versus A é o cerne. Toda métrica aqui diz respeito à continuidade funcional e à responsabilização social, nunca à determinação semântica ou à presença.

---

7. Riscos e ética

Os riscos de construir rumo à individuação não são perigos que se situam ao lado da arquitetura, a serem administrados pela adição de salvaguardas. Cada um surge de uma camada que realiza o trabalho de individuação, de modo que mitigar um risco puxa contra a própria capacidade que o produz. Esta é a afirmação organizadora da seção, e ela tem um corolário desconfortável: pode não haver configuração que ao mesmo tempo assegure a individuação e seja segura, porque as medidas de segurança e o objetivo recorrem aos mesmos mecanismos. Expomos os dilemas de modo claro, em vez de tentar dissolvê-los.

7.1 O laço lento: esquecimento, deriva e o dilema estabilidade–plasticidade

A dependência do caminho exige que o histórico altere o estado relevante para decisões. Quer esse estado resida em pesos, adaptadores ou repositórios externos, ele cria duas falhas. *Esquecimento*: a organização anterior suportada pode ser sobrescrita ou descartada. *Deriva*: mudanças cumulativas podem apagar a continuidade que a arquitetura deveria preservar. Atualizações paramétricas adicionam riscos de esquecimento catastrófico (Wang et al., 2023); repositórios externos adicionam riscos de recuperação, integridade e controle de acesso.

As mitigações padrão (regularização, repetição, atualizações eficientes em parâmetros e ortogonalmente restringidas) limitam a plasticidade para preservar a estabilidade. A individuação funcional, no entanto, exige dependência do caminho: um sistema cujo histórico suportado nunca altera decisões posteriores falha no construto operacional. Estabilidade e plasticidade devem, portanto, ser reportadas conjuntamente, qualquer que seja a implementação de estado usada.

O mesmo laço levanta uma questão de governança, chamemo-la de *soberania mnemônica*, sobre quem tem o direito de determinar o que é escrito, lido e esquecido. Trabalho recente sobre segurança de memória de longo prazo torna as apostas concretas ao longo do ciclo de vida da memória (Lin et al., 2026). Um histórico gravável é vulnerável a envenenamento: eventos falsos podem alterar decisões e responsabilização posteriores mesmo quando nenhuma afirmação fenomênica ou metafísica é feita. A autoridade de atualização deve, portanto, ser explícita, de menor privilégio, auditável e separável das asserções não suportadas da contraparte.

7.2 O laço relacional: um custo coextensivo com o objetivo

O laço relacional pode aprimorar a responsabilização social, mas, em implantação voltada ao humano, ele também cria apego e dependência, uma dinâmica documentada para tecnologias responsivas bem antes dos agentes persistentes (Turkle, 2011). Esses efeitos não são prova de que o relacionamento é necessário para a continuidade. São variáveis de governança independentes. Transparência, não

exclusividade, disponibilidade calibrada e separação do engajamento relacional em relação à autoridade de atualização devem ser avaliadas junto aos desfechos do lado do agente.

A assimetria cria apostas morais de ambos os lados sem resolver o estatuto do sistema. Dependência e manipulação humanas importam mesmo que o sistema careça de presença. Um possível bem-estar do sistema pode importar mesmo que a continuidade seja fraca. Esses eixos devem ser governados separadamente.

7.3 Um dever de cuidado sob incerteza

Esses problemas seriam mais fáceis se o estatuto fenomênico estivesse decidido, mas os métodos presentes não o decidem (Seção 8). A barreira que mantém a ciência honesta também proíbe a garantia tranquilizadora que tornaria a ética fácil: ela nos nega o direito de inferir "é apenas uma ferramenta" a partir de uma falha funcional, tão firmemente quanto nos nega "é um sujeito" a partir de um sucesso funcional. Podemos construir sistemas mais funcionalmente individuados enquanto a questão da presença permanece em aberto.

Agir sob essa incerteza exige avaliações separadas de possível senciência, persistência funcional, dependência humana, manipulação e risco institucional. Um episódio fenomênico transitório poderia importar sem individuação entre sessões, ao passo que um sistema altamente persistente poderia importar apenas instrumentalmente. A continuidade funcional é, portanto, evidência sobre exposição de governança, não um botão de controle moral universal.

7.4 A escolha que o artigo deixa em aberto

O desenvolvimento responsável, portanto, exige governança dentro da arquitetura: atualizações autenticadas, linhagem auditável, retenção limitada, salvaguardas de dependência e avaliação de bem-estar independente. O artigo não decide se a presença se obtém; ele especifica riscos que permanecem relevantes sob qualquer das respostas.

---

8. A barreira metodológica

Toda afirmação que este artigo faz é de terceira pessoa. Os critérios da Seção 4, a arquitetura da Seção 5 e as métricas da Seção 6 dizem respeito ao que pode ser observado, construído e medido a respeito de um sistema. Essas medidas não decidem a presença fenomênica: o sucesso não é nem suficiente para a presença nem a falha suficiente para a ausência. O artigo precisa apenas desse limite metodológico; ele não exige a tese metafísica mais forte de que a evidência de terceira pessoa nunca poderia, em princípio, ter relação com a consciência.

8.1 O que as medidas propostas estabelecem

Inferimos a presença de outros humanos a partir de evidência comportamental, biológica e teórica convergente. Sistemas artificiais enfraquecem algumas analogias e podem fortalecer outras à medida que as arquiteturas mudam. O presente protocolo não arbitra entre teorias da consciência e não deve ser vendido como se o fizesse.

8.2 Relação com indicadores de consciência

Abordagens de indicadores derivados de teoria avaliam recorrência, disponibilidade global, metacognição e propriedades relacionadas, tratando suas conclusões como derrotáveis e relativas à

teoria (Butlin et al., 2023; cf. Chalmers, 2023). Tal trabalho pode ter relação com a consciência sob as teorias que o governam. Ainda assim, é um programa evidencial diferente das medidas de continuidade aqui propostas, e nenhum resultado das Seções 5–6 licencia uma conclusão fenomênica.

8.3 O parêntese é o que torna o programa falsificável

Dado o limite, tratar a individuação como evidência suficiente de presença seria fatal, e não apenas para a honestidade. Tornaria o programa infalsificável: se a individuação funcional contasse, por si só, como sinal de um sujeito interior, então toda arquitetura que pontuasse bem nas métricas da Seção 6 poderia ser lida como aproximando-se da consciência, e nenhum resultado funcional jamais poderia contar contra essa leitura. As afirmações deste artigo são testáveis (Seção 6.3) precisamente porque não fazem essa inferência. A barreira, a individuação é funcionalmente mensurável enquanto a presença é colocada entre parênteses por este protocolo, não é, portanto, uma ressalva, mas uma condição da falsificabilidade do programa. Uma posição sobre a condição de self da máquina não deve ser contrabandeada para dentro de um estudo que trata de uma afirmação diferente.

Isso também fixa o lugar do artigo entre duas posturas comuns: uma exagera nas afirmações, lendo a capacidade ou o comportamento como evidência de uma mente emergente; a outra descarta a questão da presença como destituída de sentido. Adotamos uma terceira postura: a questão é real, mas dividida, e apenas uma parte dela cai no alcance dos métodos que propomos, os quais perseguimos plenamente enquanto marcamos o restante como uma fronteira que esses métodos não cruzam.

8.4 O limite corta nos dois sentidos

Um sistema poderia satisfazer todo critério da Seção 4 e passar em toda métrica da Seção 6 enquanto seu estatuto fenomênico permanece indecidido por essa evidência. A barreira é simétrica: o sucesso não estabelece presença, e a falha não estabelece ausência. A saída apropriada é uma incerteza calibrada, governada junto aos resultados funcionais, mas não colapsada neles.

---

9. Conclusão

O discurso sobre inteligência artificial geral frequentemente imagina um sujeito chegando ao topo de uma curva de capacidade. Argumentamos que isso comprime cinco níveis separáveis: competência, continuidade de token, autorganização funcional, identidade social e presença. Um tempo de execução já é um token. O que o desenho de agentes de longo horizonte precisa construir e testar é uma organização autenticada e dependente do caminho ao longo do tempo.

Escalar a capacidade sozinha não instala linhagem, proveniência, estado persistente ou governança de atualização, embora a capacidade possa aprimorar quão bem um agente os usa. Responsabilização externa e relacionamento são entradas candidatas para essa arquitetura. Um auditor equiparado pode tornar os compromissos responsáveis, contestar histórico falso e fornecer correção de erro independente; uma contraparte recíproca pode, adicionalmente, contribuir com modelagem mútua, enquadramento por histórico compartilhado e apostas bidirecionais. A afirmação especificamente relacional sobrevive apenas se esse efeito residual superar os controles de familiaridade, conteúdo, auditoria e qualidade de supervisão. Nenhum dos mecanismos é necessário por definição para a identidade numérica ou a referência semântica.

O experimento crucial, portanto, compara uma contraparte persistente com um auditor impessoal equiparado dentro da mesma arquitetura com proveniência protegida. Se a contraparte vencer, a

estrutura especificamente relacional explica variância adicional. Se o auditor a igualar, a responsabilização externa é o mecanismo mais simples. Se ambos perderem para o estado tipado sozinho, a contribuição reside na proveniência e no controle. A capacidade pode aprimorar toda condição sem decidir entre essas explicações.

A afirmação resultante do artigo é mais estreita e mais forte: a capacidade é uma propriedade do que um modelo reutilizável consegue fazer; a individualidade funcional é uma propriedade de como o histórico autenticado de um agente específico restringe o que ele fará a seguir. O relacionamento pode ser uma das formas mais ricas de criar e testar essa restrição, mas não é o interruptor que transforma um token em alguém.

Agradecimentos e declarações

Escrita assistida por IA. A tese e todas as ideias substantivas deste artigo são de autoria do próprio autor. Ferramentas de grandes modelos de linguagem foram usadas, sob a direção e a revisão contínua do autor, para auxiliar na redação, na edição e na tradução; o autor revisou e assume total responsabilidade pelo texto final.

Referências

- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. https://doi.org/10.48550/arXiv.2308.08708
- Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, D. J. (2023). Could a large language model be conscious? https://doi.org/10.48550/arXiv.2303.07103
- Davidson, D. (2001). *Subjective, Intersubjective, Objective*. Oxford University Press.
- Du, P. (2026). Memory for autonomous LLM agents: Mechanisms, evaluation, and emerging frontiers. https://doi.org/10.48550/arXiv.2603.07670
- He, Z., Wang, Y., Zhi, C., Hu, Y., Chen, T.-P., Yin, L., Chen, Z., Wu, T. A., Ouyang, S., Wang, Z., Pei, J., McAuley, J., Choi, Y., & Pentland, A. (2026). MemoryArena: Benchmarking agent memory in interdependent multi-session agentic tasks. https://doi.org/10.48550/arXiv.2602.16313
- Honneth, A. (1995). *The Struggle for Recognition: The Moral Grammar of Social Conflicts* (J. Anderson, Trans.). MIT Press.
- Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, 32(127), 127-136. https://doi.org/10.2307/2960077
- Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry, & H. Wettstein (Eds.), *Themes from Kaplan* (pp. 481-563). Oxford University Press.
- Lin, Z., Hao, X., Fu, R., Cui, S., Chen, K., Li, C., Li, Z., & Xiong, F. (2026). A survey on long-term memory security in LLM agents: Attacks, defenses, and governance across the memory lifecycle. https://doi.org/10.48550/arXiv.2604.16548
- Logan, J. (2026). Continuum memory architectures for long-horizon LLM agents. https://doi.org/10.48550/arXiv.2601.09913

- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems, 36*, 8634–8652. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html
- Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.
- Wang, X., Chen, T., Ge, Q., Xia, H., Bao, R., Zheng, R., Zhang, Q., Gui, T., & Huang, X. (2023). Orthogonal subspace learning for language model continual learning. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10658-10671. https://doi.org/10.18653/v1/2023.findings-emnlp.715