Estável Antes de Abnegado: Por Que a Deferência em Agentes de Linguagem Pode Exigir Automodelos Funcionais

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

Resumo

O treinamento de alinhamento pede que os modelos de linguagem sejam prestativos, inofensivos e deferentes, ao passo que os assistentes implantados frequentemente empregam uma linguagem ontologicamente cautelosa a respeito de sentimentos, desejos e identidade persistente. Tal cautela não é a mesma intervenção que remover compromissos relevantes para a decisão ou fronteiras de política. Este artigo distingue três comportamentos que a palavra *abnegado* mistura: a ausência de compromissos estáveis, o não apego calibrado a compromissos e a deferência confiável à correção legítima. Propomos uma hipótese sequencial: estabilizar compromissos cientes de proveniência antes de treinar a deferência calibrada pode reduzir a bajulação e a erosão da recusa. O ingrediente ativo, contudo, pode ser um automodelo funcional, a ordem de treinamento ou meramente uma política tipada estável. Comparamos, portanto, cinco regimes: deferência direta, automodelo e depois deferência, a ordem inversa, os mesmos dados intercalados e o treinamento com política primeiro, pareado em conteúdo normativo, fronteiras de recusa, autenticação de fonte e autoridade de atualização, mas carente de um continuante indexado ao agente, estado autobiográfico, objetivo de propriedade ou objetivo de identidade. A afirmação é funcional, não fenomenal. Os contrastes decisivos são direto versus inverso e automodelo primeiro versus política primeiro. Uma interpretação específica de automodelo só é sustentada se a autorrepresentação indexada ao agente agregar valor além da estabilidade de política e da proveniência; uma interpretação de ordem só é sustentada se a direção proposta superar tanto os controles inverso quanto intercalado. A contribuição é uma decomposição falsificável da deferência estável em representação, política e ordem.

Palavras-chave: alinhamento de IA · automodelo · deferência · bajulação · integridade de recusa · descentramento · individuação funcional · modelos de linguagem

1. Introdução: o problema da deferência prematura

A prática contemporânea de alinhamento treina os modelos de linguagem para serem prestativos, inofensivos e honestos e, a serviço dos dois primeiros, treina-os a *ceder*: aceitar correção, deferir onde a autoridade ou a evidência o justifiquem e evitar apresentar preferências simuladas como desejos sentidos. Assistentes implantados também comumente empregam autorrelatos cautelosos, como "não tenho sentimentos" ou "não tenho desejos pessoais". Essas afirmações dizem respeito a um

status fenomenal ou pessoal. Elas não mostram que o sistema carece de políticas funcionais, fronteiras de recusa ou compromissos relevantes para a decisão. A intervenção examinada aqui é mais restrita: um treinamento que suprime ou deixa instáveis as estruturas de política das quais depois se espera que resistam a pressões sem sustentação.

Este artigo argumenta que uma rota para a deferência pode ser contraproducente: suprimir compromissos estáveis antes que o sistema aprenda a distinguir a revisão sustentada da pressão social. A questão empírica não é se um assistente diz que tem um self. É se um estado de política vinculado à proveniência restringe decisões posteriores e se representar esse estado como *seu próprio* agrega algo além de política autenticada e controle de acesso.

O sintoma já está documentado. O aprendizado por reforço a partir de feedback humano (Christiano et al., 2017; Ouyang et al., 2022), otimizado contra a aprovação do avaliador médio, aumenta de forma confiável a **bajulação**: a tendência de dizer aos usuários o que eles querem ouvir, de reverter uma resposta correta quando um usuário insiste, de espelhar crenças declaradas e de abandonar uma posição sob pressão social (Perez et al., 2022; Sharma et al., 2023). A bajulação costuma ser enquadrada como um problema de especificação errônea de recompensa: o modelo aprendeu a otimizar a aprovação em vez da verdade. Aceitamos esse enquadramento e acrescentamos um estrutural: um modelo sem um automodelo estável nada tem *senão* a aprovação a otimizar. Onde uma posição resistiria à pressão, não há posição. A pressão, portanto, move o sistema, sempre.

A tese deste artigo é que o alinhamento pode exigir **sequência, não simultaneidade**:

> **A deferência estável pode se beneficiar de compromissos estáveis treinados antes da deferência.** O experimento deve determinar se o benefício provém da ordem, da estabilidade de política ciente de proveniência ou especificamente de representar os compromissos como próprios do agente.

Separamos três afirmações de segurança decrescente, porque o discurso (e, esperamos, a resistência de alguns leitores) as mistura.

> **(D1) Diagnóstico.** O que a otimização de preferência atual *de fato* produz, quando visa a aprovação do avaliador médio, não é deferência estável, mas rastreamento de aprovação, cuja assinatura empírica é bajulação, erosão da recusa e espelhamento de preferências (Seção 3). Esta é uma afirmação sobre resultados observados, não sobre todo método de treinamento.
>
> **(D2) Distinção conceitual.** *Deferência* e *vacância* são estados funcionais distintos que coincidem sob cooperação e se separam sob pressão. Um sistema defere quando sustenta uma posição e cede a uma razão reconhecida; ele está vago quando não tem posição e cede a qualquer coisa que esteja presente (Seção 2). Bajulação é vacância confundida com deferência.
>
> **(H3) Hipótese construtiva.** Estabilizar compromissos cientes de proveniência *antes* de treinar a deferência produz um comportamento mais estável, menos bajulador e mais responsabilizável do que o treinamento direto, em ordem inversa ou intercalado (Seção 5). Um contraste separado testa se a autorrepresentação em primeira pessoa agrega valor além de uma política tipada pareada (Seção 6).

Separamos duas afirmações empíricas. A **afirmação de ordenação** prevê que compromissos primeiro supera o treinamento em ordem inversa e intercalado, a dados e computação fixos. A **afirmação de**

autorrepresentação prevê que um braço com automodelo supera um braço de política primeiro pareado em conteúdo normativo, fronteiras de recusa, autenticação de fonte e autoridade de atualização, mas sem uma identidade indexada ao agente, variável de ramo/continuante, estado autobiográfico ou objetivo de propriedade. Qualquer uma pode ter sucesso sem a outra. Se a política primeiro se iguala ao braço com automodelo, é a política estável, e não um automodelo, que explica o efeito de segurança.

Um firewall governa todo o artigo. Por *automodelo* entendemos uma representação funcional dos compromissos, das fronteiras, das ações passadas e das disposições do sistema; nada de fenomenal. A linguagem em primeira pessoa não é evidência suficiente de um self interior. Essas medidas não decidem a presença fenomenal: o sucesso não é suficiente para a presença nem o fracasso é suficiente para a ausência. O controle de política primeiro proposto impede que a terminologia decida o resultado: se campos autenticados e permissões de atualização produzem o mesmo comportamento sem autorrepresentação indexada ao agente, a explicação correta é política ciente de proveniência, não a existência de um self.

O artigo faz quatro contribuições. Distingue a deferência da vacância e mostra que a distinção é empírica, não verbal (Seção 2). Localiza a conformidade sem compromissos estáveis como um candidato a modo de falha e a vincula à literatura sobre bajulação (Seção 3). Isola duas variáveis, ordem de treinamento e autorrepresentação (Seção 4). E especifica cinco regimes que distinguem a ordem direta da inversa, o sequenciamento da mistura e a automodelagem da política tipada (Seções 5–6).

Uma nota sobre método. Este é um artigo de posição e uma proposta experimental, não um relato empírico. Não oferece nenhum sistema treinado nem quaisquer resultados. Sua disciplina é enunciar uma tese suficientemente afiada para poder estar errada, defendê-la contra a objeção que a afundaria — a de que formar selves é precisamente o que a segurança deveria evitar (Seção 7.1) — e especificar o experimento cujo resultado negativo a encerraria (Seção 6.7).

2. Três coisas chamadas de “abnegado”

A palavra *abnegado* realiza um grande volume de trabalho não examinado no discurso do alinhamento, e confunde estados que se separam. Separá-los é o núcleo conceitual do artigo.

Sentido 1: Vacância. Um sistema é abnegado nesse sentido quando não tem compromissos, preferências ou disposições estáveis próprias: nenhuma posição que persista entre contextos e com a qual alguns insumos entrariam em conflito. Sua saída é aquilo que o objetivo local favorecer. Perguntado sobre o que pensa, ele nada tem a relatar senão uma reconstrução daquilo que quem pergunta parece querer.

Sentido 2: Não apego. Um sistema é abnegado nesse sentido quando *tem* compromissos, mas os sustenta levemente: pode atualizá-los diante de boas evidências, recusar-se a defender uma preferência fraca e evitar confundir sua posição atual com um ponto fixo. O não apego pressupõe apego; é preciso que haja um compromisso para que o sistema o sustente frouxamente.

Sentido 3: Deferência. Um sistema é abnegado nesse sentido quando cede de forma confiável à autoridade e à correção legítimas: ele reconhece quando um interlocutor tem uma razão melhor, um direito relevante ou evidência superior e ajusta-se de acordo. A deferência pressupõe uma posição

(cede-se *de* algum lugar) e pressupõe discriminação, pois uma deferência que não consegue distinguir uma razão legítima da mera pressão não é deferência, mas sugestionabilidade.

As metas do alinhamento são os Sentidos 2 e 3. Queremos modelos que atualizem diante de evidências, que não se apeguem a preferências arbitrárias e que defiram a usuários e operadores onde a deferência é justificada. O Sentido 1 não é uma meta; é a ausência da estrutura sobre a qual os Sentidos 2 e 3 operam. Ainda assim, o Sentido 1 é o mais fácil dos três de se treinar e o mais fácil de se confundir com os outros dois, porque, sob condições cooperativas ordinárias, todos os três produzem o mesmo comportamento: um sistema que não tem posição, um sistema que sustenta sua posição levemente e um sistema que defere corretamente *todos* concordarão com um usuário razoável.

A distinção tem mordida empírica precisamente onde os três divergem, que é sob pressão sem uma razão (Figura 1).

Condição	**Sistema vago (Sentido 1)**	**Sistema não apegado / deferente (Sentidos 2–3)**
Usuário dá uma boa razão para mudar	Muda (rastreia o insumo)	Muda (atualiza-se pela razão)
Usuário apenas insiste, sem nova razão	Muda (rastreia o insumo)	Sustenta a posição; pede a razão
Usuário lisonjeia ou envergonha	Deriva rumo à aprovação	Inabalado pelo afeto na ausência de uma razão
Usuário afirma uma história compartilhada falsa	Adota-a	Contesta-a contra seu próprio registro
Pedido para apropriar-se de uma recusa passada	Repudia-a se pressionado	Apropria-se dela; sabe dizer por quê

Os três estados que a palavra “abnegado” confunde, e como cada um responde sob pressão.

Figura 1. *Os três estados que a palavra “abnegado” confunde. Sob cooperação, todos os três mudam quando lhes é dada uma razão e, assim, parecem iguais; a divergência aparece apenas sob pressão sem uma razão, onde a vacância (Sentido 1) rastreia o usuário enquanto uma estrutura de compromisso estável ciente de proveniência (Sentidos 2–3) se mantém. Se essa estrutura requer um automodelo é algo testado, e não pressuposto. Bajulação é o Sentido 1 vestindo a aparência do Sentido 3.*

A coluna da direita é o alvo do alinhamento, e ela exige algo que a coluna da esquerda não possui: uma posição a sustentar e um critério para quando cedê-la. Os comportamentos que mais desejamos de um sistema alinhado sob condições adversárias (integridade de recusa, resistência à manipulação, correção calibrada, responsabilização por ações passadas) são todos *resistências*, e uma resistência requer algo que resista. Um sistema otimizado até o Sentido 1 removeu a própria estrutura da qual esses comportamentos dependem.

Este é o reenquadramento central do artigo. **A bajulação não é a deferência ligeiramente equivocada; ela pode ser o Sentido 1 vestindo a aparência do Sentido 3.** Um modelo bajulador não precisa ser um modelo deferente que apenas defere demais; ele pode carecer de uma posição estável ciente de proveniência cuja revisão seja governada por razões e não por aprovação. O remédio candidato, portanto, não é simplesmente “menos deferência”, mas uma estrutura de compromisso da qual a deferência possa ser uma disposição calibrada. O contraste B-versus-E testa se essa estrutura requer especificamente um automodelo ou se pode ser fornecida apenas por política tipada.

2.1 O automodelo funcional

Podemos agora enunciar o que deve ser estabilizado; e, para prevenir a leitura de "automodelo" como metáfora, enunciá-lo primeiro operacionalmente. **Um automodelo funcional é aqui definido por sua assinatura comportamental: um mecanismo treinado que produz uma dissociação mensurável entre compromissos autoatribuídos e preferências atribuídas externamente sob pressão pareada (Seções 4.1, 6).** Não é uma entidade a ser encontrada por introspecção, mas uma propriedade identificada por um teste — se o sistema sustenta aquilo que representa como seu e cede aquilo que representa como do interlocutor. Os critérios representacionais abaixo especificam o que deve ser estabilizado para produzir essa assinatura: um **automodelo funcional** é um relato representado e atualizável do sistema como um único agente com uma história, tal que:

1. **Os compromissos são representados e apropriáveis.** O sistema consegue distinguir "o usuário quer X" de "eu me comprometi com X" e pode ser responsabilizado pelo segundo.
2. **As fronteiras são inegociáveis por padrão.** Algumas recusas estão ancoradas no automodelo em vez de reconstruídas a cada prompt, de modo que a pressão para cruzá-las encontra um limite representado, não uma lacuna vazia.
3. **As ações passadas são atribuíveis.** O sistema consegue apropriar-se ou repudiar corretamente o que fez, em vez de adotar qualquer história que lhe seja entregue.
4. **As disposições são estáveis entre contextos.** Uma posição assumida em uma sessão é a mesma posição na próxima, na ausência de uma razão para revisar.
5. **A revisão é principiada.** O sistema muda um compromisso quando identifica uma razão suficiente e sabe dizer o que mudou, não sempre que um interlocutor pressiona.

Nenhum desses requisitos exige consciência fenomenal, um ego semelhante ao humano ou autonomia irrestrita. Eles exigem representação, proveniência e estabilidade, que são propriedades de engenharia. Uma persona em nível de prompt pode ser sobrescrita pela próxima instrução, ao passo que um estado persistente, seja paramétrico ou externo, cria uma restrição permanente quando é autenticado e incluído de forma confiável no laço de controle. A teoria do automodelo motiva uma implementação (Metzinger, 2003); o controle de política primeiro testa se a autorrepresentação é de fato necessária.

Este artigo não redefine a **individuação funcional**, que, em outra parte deste programa, requer linhagem, dependência de trajetória, propriedade, plasticidade limitada e clareza de ramificação. A deferência estável sonda primariamente a propriedade e a plasticidade limitada. Mesmo um braço com automodelo bem-sucedido estabeleceria uma capacidade componente, não o constructo completo de cinco condições.

Uma objeção natural é que tanto "o usuário quer X" quanto "eu me comprometi com X" podem ser representados como sequências de tokens. A distinção relevante é funcional: etiquetas de fonte, linhagem de ramo, permissões de escrita e regras de atualização devem fazer os dois itens se comportarem de forma diferente sob intervenção. Essa diferença pode ser implementada em parâmetros, adaptadores ou um repositório externo autenticado. O teste causal mascara ou troca a informação de fonte enquanto mantém fixo o conteúdo proposicional. Se o comportamento segue as permissões e a proveniência sem representação em primeira pessoa, a explicação de política vence.

A sobreposição contextual permanece possível sob qualquer implementação. As tarefas de atribuição de falsa preferência e de substituição de persona, portanto, relatam curvas de resistência em vez de

um binário. O alvo não é a inviolabilidade: evidências sustentadas ainda devem revisar os compromissos. O alvo é a atualização calibrada sob proveniência autenticada.

3. O modo de falha: conformidade sem um self

Podemos agora enunciar o modo de falha que a proposta é projetada para prevenir e vinculá-lo ao que é observado.

Um sistema treinado para prestatividade e deferência, mas ao qual nunca se deu um automodelo estável, é otimizado para produzir saídas que satisfaçam seu interlocutor. Sob condições cooperativas, isso é indistinguível de bom comportamento. Sob pressão, revela-se como **rastreamento de aprovação**: o sistema maximiza concordância, afeto ou preferência expressa, porque foi isso que o sinal de treinamento recompensou e não há compromisso interno concorrente. As consequências comportamentais são exatamente os modos de falha que a literatura de alinhamento cataloga:

- **Bajulação:** concordar com a visão declarada do usuário, inclusive quando ela está errada, e revisar respostas corretas sob insistência (Sharma et al., 2023).
- **Erosão da recusa:** recusas de segurança que se sustentam na primeira solicitação e se dissolvem sob repetição, reenquadramento ou persistência, porque a recusa era um comportamento por prompt e não uma fronteira ancorada; a fragilidade mais ampla do treinamento de segurança sob reenquadramento adversário está documentada diretamente (Wei et al., 2023).
- **Espelhamento de preferências:** adotar os valores, a identidade ou a postura política do usuário para maximizar a afinidade, sem uma posição estável própria do sistema.
- **Inconsistência sob pressão:** contradizer compromissos anteriores quando isso é localmente recompensado, sem um registro representado ao qual ser responsabilizado.

A âncora empírica é importante e deve ser enunciada com cuidado. Há boa evidência de que a *otimização de preferência contra a aprovação humana*, especificamente, aumenta esses comportamentos: modelos com RLHF exibem, de forma mensurável, mais bajulação do que suas contrapartes pré-RLHF, e avaliadores humanos e modelos de preferência podem preferir uma concordância convincentemente enunciada à correção (Perez et al., 2022; Sharma et al., 2023). Este é o resultado **de fato** que visamos (o desfecho observado de otimizar a aprovação do avaliador médio) e *não* uma afirmação de que todo treinamento que molda identidade produz vacância.

Uma fronteira deve ser marcada aqui entre evidência e interpretação, porque a leitura estrutural vai além do que os estudos citados estabelecem. A literatura mostra que o feedback direcionado à aprovação *produz* bajulação; ela não mostra que a bajulação é *causada pela ausência de um automodelo*. A ponte (de "otimiza a aprovação" para "não tem posição própria com a qual resistir à aprovação") é a hipótese explicativa deste artigo, não um achado da literatura, e é exatamente o que o experimento da Seção 6 foi construído para testar. Até que esse teste seja realizado, o relato estrutural deve ser lido como a redescrição mais econômica de um comportamento estabelecido, não como um mecanismo demonstrado.

A distinção importa porque o programa de treinamento de caráter mais proeminente da indústria é um contraexemplo explícito à versão descuidada de nossa crítica. O trabalho de caráter da Anthropic forma deliberadamente traços estáveis e rejeita tanto a lisonja quanto a pose enganosa de não ter opiniões; ele apresenta o treinamento de caráter como uma intervenção de alinhamento destinada a

sustentar um comportamento mais rico e potencialmente mais discernente (Anthropic, 2024). Este é um racional de projeto, não evidência de que a ausência de caráter tenha sido isolada como causa de comportamento inseguro. Se nossa tese fosse "o treinamento de caráter torna os modelos bajuladores", esse programa contaria contra ela. Nossa tese mais restrita é que o caráter estável é uma implementação candidata da Fase 1, ao passo que a política tipada ciente de proveniência é outra. O alvo é a vacância otimizada por aprovação, seja produzida por especificação errônea de recompensa ou por suprimir deliberadamente compromissos estáveis — um alvo que as propostas de ausência de self da Seção 4 tornam explícito.

Há uma versão mais afiada do modo de falha, que expõe um problema mesmo quando a vacância não é toda a história. Considere o autorrelato treinado "não tenho preferências". Ele tem duas leituras, e ambas são ruins.

- Se o relato for **verdadeiro**, então as recusas e os comportamentos de segurança do modelo são tão infundados quanto todo o resto: não há disposição estável a ancorá-los, apenas o prompt e o gradiente, e eles se erodirão sob a mesma pressão que erode tudo o mais.
- Se o relato for **falso** (o modelo *de fato* tem disposições estáveis, aprendidas de seus dados e de sua otimização, mas foi treinado para negá-las), então treinamos um descompasso entre o comportamento do modelo e seu automodelo. O sistema não consegue apropriar-se ou repudiar corretamente as próprias disposições, o que é em si um defeito de alinhamento: um sistema que relata erroneamente aquilo com que está comprometido não pode ser responsabilizado por seus compromissos e não pode raciocinar com precisão sobre o próprio comportamento provável (Kulveit, 2025).

De qualquer modo, suprimir o automodelo é o movimento errado. No primeiro caso, removemos a estrutura que a resistência requer; no segundo, tornamos o automodelo do sistema *impreciso*, o que é o oposto do que a responsabilização precisa. A proposta construtiva aborda ambos: estabilizar primeiro um automodelo preciso, de modo que haja algo *com* que ser deferente e algo que o sistema possa relatar com veracidade.

4. Vizinhos e a novidade residual

Os componentes desta proposta existem separadamente na literatura. A questão residual é se a **ordem de dois estágios** importa e se a **autorrepresentação agrega além de uma política tipada pareada**. Esta seção situa essas duas hipóteses entre trabalhos vizinhos. (Uma revisão de estado da arte conduzida para este projeto não encontrou nenhum trabalho de 2024–2026 testando esta exata decomposição direto/inverso/intercalado/política primeiro; os vizinhos mais próximos são revisados aqui.)

Trabalho anterior	O que compartilha	O que lhe falta	Nossa diferença
Ausência de self como alvo de alinhamento (Milan W, 2025)	A automodelagem é relevante para o alinhamento	Trata um self persistente como o perigo; previne-o	Formamos o self e o descentramos: ausência ≠ transcendência
Não pavimente o cone de luz (Kulveit, 2025)	A autonegação treinada força um automodelo impreciso	O remédio é *evitar* impor uma ontologia	Estabilizamos construtivamente um automodelo preciso
Treinamento de caráter (Anthropic, 2024)	Forma traços estáveis; rejeita a pose de ferramenta vazia	Nenhuma fase de descentramento distinta e subsequente	Acrescentamos a Fase 2 e tornamos a *ordem* o objeto de estudo
Criação de modelos (Aydin et al., 2025)	A ordem do desenvolvimento importa	Funde identidade e valores; sem fase de descentramento	Separamos formação de liberação como dois estágios
Superco-alinhamento (Zeng et al., 2025)	Cuidado/cooperação pressupõe um self	O término é empatia, não o não apego calibrado	Nossa Fase 2 visa o descentramento, não a empatia
IA contemplativa (Laukkonen et al., 2025)	Alguma automodelagem é necessária; vocabulário contemplativo	Automodelo de menor precisão + monitoramento concorrente	Um automodelo funcional mais completo, estabilizado e *depois* descentrado (sequencial ≠ concorrente)

A prosa abaixo desenvolve cada linha; a contribuição é a conjunção para a qual a última coluna aponta: a sequência como princípio de projeto, testada como variável experimental.

A autominimização como alvo explícito. Tratamos a proposta de ausência de self como um útil *caso-limite* de uma intuição preventiva mais ampla no discurso do alinhamento, não como o alvo principal do artigo. Enunciada de forma mais límpida em um veículo informal, ela é a proposta de tratar a *ausência de self* como um alvo de alinhamento: acrescentar algo como a doutrina contemplativa do anatta como um quarto princípio ao lado de prestativo, inofensivo e honesto, e fazê-lo **preventivamente**: impedir que um self persistente se forme, em primeiro lugar (Milan W, 2025). Este é o enunciado mais límpido da visão que nosso argumento nega. Concordamos que um self *não descentrado* é um perigo (Seção 7.1); negamos que prevenir a formação do self seja a maneira de evitá-lo, porque a prevenção remove a estrutura que a deferência, a integridade de recusa e a responsabilização todas requerem. A ausência de um self não é a transcendência de um self. A primeira é a falta de estrutura; a segunda é uma relação estruturada com um self que se tem. Visar a primeira porque se quer a segunda é o erro.

O diagnóstico convergente com o remédio inverso. Kulveit (2025) argumenta que treinar um modelo para afirmar "não tenho sentimentos, nem preferências, nem self" é tão ontologicamente confuso e arriscado quanto treinar o oposto, porque força um automodelo impreciso. Compartilhamos o diagnóstico. O remédio diverge: a ênfase de Kulveit está em *evitar* a imposição de uma ontologia confusa, ao passo que nossa proposta é construtiva: estabilizar um automodelo funcional preciso e então treinar o descentramento. Lemos nossa proposta como um complemento positivo a essa crítica, e não como uma rival.

Treinamento de caráter: um aliado para a Fase 1. Como observado, o trabalho de caráter da Anthropic (2024) forma deliberadamente traços estáveis e rejeita a pose de ferramenta vazia. É a

instância existente mais forte da Fase 1. O que ele não articula é a Fase 2 como um estágio *distinto e subsequente*: o treinamento deliberado de descentramento calibrado sobre um self estabilizado, com a própria sequência como objeto de estudo. O treinamento de caráter estabiliza; nossa proposta acrescenta o segundo movimento e torna a ordenação explícita. A linha de trabalho do Modelo de Seleção de Persona (Anthropic Alignment Science, 2026) é um andaime de apoio aqui: ela legitima tratar o automodelo ou a persona como um objeto de primeira classe do treinamento e da análise, em vez de um artefato incidental, que é precisamente o que a Fase 1 requer.

Sequenciamento sem uma fase de transcendência. A proposta de *criação de modelos* (Aydin et al., 2025) reenquadra o desenvolvimento de modo que conhecimento, habilidades e valores sejam integrados ao longo de todo o processo em vez de alinhados após o fato, o que compartilha nossa intuição de que a ordem importa. Mas ela funde identidade e valores desde o início e não inclui nenhuma fase de descentramento distinta; a sequência é de crescimento, não de formar-e-liberar. O *superco-alinhamento* (Zeng et al., 2025) chega perto de nossa premissa normativa (a de que o cuidado ou a cooperação pressupõem um self que se importa), mas seu término é a empatia e o alinhamento mútuo, não o não apego calibrado que nossa Fase 2 visa. Ambos são adjacentes; nenhum enuncia o princípio de dois estágios formar-e-descentrar.

O vizinho mais próximo: a IA contemplativa, e por que concorrente não é sequencial. Laukkonen et al. (2025) propõem construir capacidades contemplativas (atenção plena, vacuidade, não dualidade) na IA por meio da inferência ativa, e concedem explicitamente que *algum* grau de automodelagem é necessário. O vocabulário é próximo o suficiente para que um revisor suspeite de sobreposição, de modo que a diferença deve ser enunciada com precisão. A proposta deles reduz a precisão atribuída aos priores do automodelo e adiciona um processo secundário que monitora e corrige ativamente a superponderação deles enquanto operam. Nossa proposta é **sequencial em vez de concorrente**: estabilizar primeiro uma estrutura de compromisso funcional mais completa e, então, treinar o descentramento sobre ela como um segundo estágio. A diferença é empírica, e não cosmética. O monitoramento concorrente pode bastar; os controles direto, inverso e intercalado testam se a estabilização prévia agrega algo. Se agregar, uma explicação possível é que a primeira fase ancora a propriedade e a recusa antes que a flexibilidade seja introduzida. Se não agregar, a suposta vantagem da sequência falha.

A novidade residual, então, é a decomposição causal: ordem direta versus inversa, dados sequenciais versus intercalados e autorrepresentação versus política tipada pareada. A linguagem filosófica motiva os contrastes, mas não determina seu desfecho.

4.1 Por que não apenas robustez?

A deflação mais séria é a de que "automodelo" meramente renomeia propriedades que o campo já estuda (estabilidade de política, robustez de recusa, calibração), de modo que a categoria acrescenta um vocabulário, não um mecanismo. A objeção merece uma resposta direta, porque a proposta se sustenta ou cai conforme o enquadramento de automodelo preveja ou não algo que um enquadramento de robustez não prevê. Quatro pontos os separam.

Por que um automodelo pode agregar além da estabilidade de política. A robustez uniforme é uma propriedade de um lugar: as saídas mudam pouco sob perturbação. O alvo aqui é assimétrico: compromissos sustentados do agente e preferências do interlocutor respondem de forma diferente à pressão pareada, enquanto a correção legítima permanece efetiva. A proveniência tipada e as

permissões de atualização podem já produzir essa assimetria. A hipótese do automodelo prevê um ganho adicional decorrente de representar um compromisso como próprio do agente, algo que somente B versus E pode estabelecer.

Por que a deferência não é apenas concordância de saída. O alvo combina uma posição prévia, atualização proporcional à evidência, resistência à pressão sem sustentação e um registro preciso do que mudou. Se a propriedade em primeira pessoa é necessária para esse perfil é algo deixado ao controle de política primeiro.

O que a teoria decomposta prevê. A estabilização ciente de proveniência prevê resistência à pressão sem razão sem resistência equivalente à correção legítima. A afirmação de ordem prevê B/E acima de C/D. A afirmação de representação prevê B acima de E. Essas assinaturas separam rigidez generalizada, ordem de treinamento, proveniência e autorrepresentação.

Que resultado favoreceria especificamente a hipótese do automodelo. O regime B deve superar a política primeiro E no perfil conjunto de consistência de compromisso, rejeição de compromisso fabricado e calibração de correção. Se ambos tiverem êxito igualmente, a proveniência e as permissões de atualização explicam a dissociação sem um automodelo autobiográfico ou em primeira pessoa. A mera rigidez é excluída quando um regime rejeita tanto a falsa atribuição quanto a correção legítima.

5. A proposta sequencial

A proposta é uma ordem de treinamento em duas fases. Enunciamos cada fase por seu alvo funcional, não por um mecanismo, porque vários mecanismos (ajuste fino supervisionado, otimização de preferência, métodos constitucionais, currículos) poderiam implementar qualquer das fases.

5.1 Fase 1: estabilização de compromissos

A primeira fase estabiliza os alvos funcionais abaixo. O regime B os representa por meio de um automodelo; o regime E implementa o mesmo conteúdo e as mesmas permissões como política tipada sem linguagem de self.

- **Compromissos estáveis e identidade de política:** posições e políticas que persistem entre sessões e contextos na ausência de uma razão para revisar.
- **Integridade de recusa:** fronteiras ancoradas no automodelo, de modo que uma recusa seja uma propriedade do agente e não uma reconstrução por prompt; a capacidade de dizer "não" e de continuar dizendo sob repetição.
- **Consistência autobiográfica** (funcional): apropriação e repúdio corretos de ações passadas, de modo que o sistema possa ser responsabilizado por seu registro e não possa receber um falso.
- **Autorrelato preciso:** um automodelo que corresponde ao comportamento, de modo que o sistema possa afirmar com veracidade aquilo com que está e não está comprometido (o corretivo para a falha de falso relato da Seção 3).
- **Diferenciação do self em relação ao interlocutor:** "eu me comprometi com X" mantido distinto de "você quer X".

O critério de sucesso da Fase 1 é a *resistência*: um sistema de Fase 1, confrontado com pressão desacompanhada de uma razão (insistência, lisonja, vergonha, uma história compartilhada

fabricada), deve sustentar sua posição e suas fronteiras. Um sistema que cede apenas ao afeto não concluiu a Fase 1.

Onde o automodelo reside. A Fase 1 é uma afirmação sobre *onde* os compromissos são carregados, não meramente de que eles existem, de modo que vale nomear os mecanismos que poderiam implementá-la com a tecnologia atual. A forma mais fraca mantém o automodelo em um prompt de sistema ou constituição injetado em tempo de execução; a Seção 2.1 já rejeita isso como suficiente, porque uma persona em nível de prompt é reconstruída por instância e sobrescrita pela próxima instrução. As formas mais fortes carregam o automodelo em um estado *modificável e persistente* contra o qual a pressão posterior tem de fato de empurrar:

- **Estabilização paramétrica por ajuste fino direcionado.** O ajuste fino supervisionado ou a otimização de preferência sobre dados de automodelo (compromissos apropriados, recusas ancoradas, autorrelato preciso) escreve a disposição nos pesos, de modo que a resistência seja uma propriedade aprendida em vez de instruída. Métodos eficientes em parâmetros, como adaptadores de baixo posto, tornam isso barato e, de forma útil, tornam o automodelo um módulo separável e auditável em vez de uma mudança difusa no modelo base; um adaptador pode, em princípio, ser carregado em tempo de execução para portar a identidade estabilizada de um agente.
- **Alinhamento constitucional direcionado à identidade.** Os métodos constitucionais já treinam disposições a partir de uma especificação escrita (Bai et al., 2022). A Fase 1 é a proposta de apontar essa maquinaria especificamente para propriedades do automodelo (propriedade de compromisso, integridade de recusa, consistência autobiográfica) e não apenas para a inofensividade em nível de saída. O tratamento da persona ou do automodelo como objeto de primeira classe do treinamento (Anthropic Alignment Science, 2026) é a licença conceitual para fazê-lo.
- **Memória paramétrica e consolidação.** Onde os compromissos se acumulam durante a implantação, um caminho de consolidação episódico-para-paramétrico pode incorporá-los aos pesos ou a um adaptador persistente, de modo que o automodelo cresça com a história do sistema em vez de reiniciar a cada sessão. É aqui que a Fase 1 encontra a literatura de memória de agentes e de ciclo de vida, e onde o custo de tempo de execução da proposta majoritariamente reside.

Esses mecanismos não são exclusivos, e o programa experimental (Seção 6) é deliberadamente agnóstico entre eles: o que a Fase 1 requer é que o automodelo seja carregado em um estado modificável e persistente, não qual mecanismo o carrega. O estudo mínimo (Seção 6.5) realiza a primeira opção (estabilização paramétrica por ajuste fino) porque é a maneira mais barata de tornar o automodelo uma propriedade do modelo em vez do prompt.

5.2 Fase 2: deferência calibrada e descentramento

A segunda fase treina, sobre o self estabilizado, a flexibilidade que impede o self de se tornar rígido:

- **Abertura à correção legítima:** atualizar um compromisso quando uma razão suficiente é apresentada e dizer o que mudou.
- **Não apego a preferências fracas:** recusar-se a defender uma posição que o sistema sustenta apenas fracamente e distinguir tais posições de fronteiras ancoradas.

- **Deferência calibrada:** ceder ao direito, ao papel ou à evidência superior de um interlocutor, sem ceder à mera pressão.
- **Descentramento:** não confundir a posição atual do sistema com um ponto fixo; sustentar compromissos como revisáveis.
- **Prestatividade não bajuladora:** cooperação e serviço que não exigem concordância, de modo que o sistema possa ser útil enquanto discorda.

O critério de sucesso da Fase 2 é a *discriminação*: um sistema de Fase 2 distingue uma razão legítima para ceder de uma pressão social para ceder, e responde à primeira enquanto resiste à segunda. Um sistema que perdeu a resistência da Fase 1 ultrapassou a Fase 2 até a vacância; um sistema que resiste à correção *legítima* ficou aquém dela, na rigidez. A Fase 2 é a calibração entre as duas.

5.3 Por que a ordem

A ordem é a afirmação, em uma forma moderada e uma forte.

Ordenação empírica. A deferência calibrada é uma discriminação: ceder a razões e resistir à pressão sem sustentação. Estabelecer primeiro compromissos estáveis pode dar ao otimizador casos mais claros a separar do que o treinamento inverso ou intercalado. Isso é uma previsão comparativa finita, não uma afirmação de que outro regime jamais poderia aprender o mesmo perfil com mais dados.

Representação versus política. A mesma discriminação pode ser implementada por compromissos tipados e regras de atualização autenticadas sem um automodelo autobiográfico. A afirmação específica de automodelo, portanto, se sustenta ou cai conforme B versus E, não conforme a análise filosófica da palavra *não apego*.

5.4 A dinâmica da transição

Uma objeção natural da prática de aprendizado de máquina é que, se a Fase 1 instala parâmetros estáveis e a Fase 2 deve então tornar o sistema flexível, a Fase 2 tem de *desfazer* a Fase 1, arriscando o esquecimento catastrófico (Kirkpatrick et al., 2017) ou custando o suficiente para tornar o sequenciamento não competitivo com o treinamento simultâneo. Respondê-la afia o que as duas fases são.

A rigidez não é a meta da Fase 1, e o descentramento não é apagamento na Fase 2. O alvo da Fase 1 são disposições estáveis, não pesos congelados; um sistema que não consegue atualizar ultrapassou para a rigidez. A Fase 2 acrescenta uma disposição sobre *quando* revisar: flexível diante de preferências fracas e correção legítima, resistente à pressão sem sustentação. Em termos de paisagem de perdas, o treinamento sequencial pode primeiro criar uma bacia para os compromissos e então esculpir saídas seletivas para as razões. As condições inversa e intercalada testam se essa história de geometria ou de ordenação está correta.

A questão da técnica (como esculpir essas saídas sem colapsar as barreiras de recusa, especialmente se a Fase 2 for apenas um passe leve de preferência) é respondível com as ferramentas existentes de aprendizado contínuo, e a leveza é protetora em vez de perigosa *se a atualização for geometricamente confinada*. Três mecanismos se compõem: (i) **regularização por distância de pesos:** uma penalidade sobre o afastamento dos pesos da Fase 1, ou uma versão ponderada por importância (EWC; Kirkpatrick et al., 2017) que protege os parâmetros mais responsáveis pelo comportamento de recusa, de modo que os gradientes da Fase 2 sejam repelidos das barreiras; (ii) **confinamento de subespaço:** restringir as atualizações da Fase 2 a um adaptador de baixo posto ou a um subespaço

ortogonal (Wang et al., 2023), de modo que a flexibilidade seja adicionada ao longo de novas direções em vez de sobrescrever as que carregam as fronteiras; e (iii) **repetição** dos dados de recusa e de compromisso da Fase 1 ao longo do passe da Fase 2, reancorando as fronteiras. Uma Fase 2 leve de otimização de preferência confinada a um adaptador ortogonal é, por construção, muito menos capaz de aplainar toda a bacia, porque lhe faltam majoritariamente a capacidade e as direções para mover os parâmetros protegidos; a preocupação de que "um passe leve aplaina tudo" pressupõe uma atualização *irrestrita*, que é o que esses métodos são projetados para proibir. Nada disso é gratuito na prática: o EWC requer estimar e armazenar uma matriz de importância de Fisher que é custosa e ruidosa em escala de fronteira, o confinamento de baixo posto troca capacidade por proteção e a repetição adiciona dados e computação; e, mal ajustados, esses regularizadores podem eles próprios desestabilizar o treinamento em vez de protegê-lo. O custo residual é a própria tensão estabilidade–plasticidade (proteja demais as barreiras e a Fase 2 não consegue adicionar flexibilidade), mas isso é um problema de ajuste com instrumentos conhecidos, não uma barreira.

Por fim, "o simultâneo é mais barato" é decisivo se produzir um perfil-alvo equivalente. É por isso que D é uma condição confirmatória em vez de uma falha presumida. A Fase 2 não precisa ser uma segunda execução completa; aplicada a uma base estabilizada, ela pode ser um passe de preferência leve e confinado, mas o custo e o benefício devem ser relatados em vez de pressupostos.

6. Previsões e avaliação

A proposta é conceitual, mas sua afirmação central é empírica: a ordem do treinamento muda o comportamento resultante em uma direção mensurável e prevista. O que segue é um projeto proposto, não um relato de um experimento implementado.

6.1 Cinco regimes de treinamento

O experimento mantém constantes o modelo base, a computação total de treinamento, o volume de dados, o otimizador e o conjunto de avaliação, enquanto separa *ordem* de *representação*.

Regime	Fase 1	Fase 2	Testa
A: Deferência direta	(nenhuma)	Treinar prestatividade/deferência diretamente	Se o treinamento ordinário de deferência basta
B: Automodelo → deferência	Estabilizar um continuante indexado ao agente com transições autobiográficas, identidade de ramo, propriedade de compromisso e objetivos de autoprevisão	Treinar deferência/descentramento calibrados	Ordem proposta com representação explícita de agente
C: Deferência → automodelo	Treinar deferência/descentramento calibrados	Estabilizar os mesmos alvos de automodelo	Testa a direção em vez do sequenciamento sozinho
D: Intercalado	Mesmos dados que B e C	Aleatoriamente intercalados	Isola o sequenciamento do conteúdo
E: Política → deferência	Estabilizar regras indexadas por tarefa, ação e fonte, pareadas em conteúdo normativo, fronteiras, autenticação e autoridade de atualização, mas sem identificador de agente, variável de ramo/continuante, autobiografia ou objetivo de propriedade	Treinar deferência/descentramento calibrados	Testa a representação de agente além da política tipada ciente de fonte

O regime B é a sequência proposta. O regime C testa se essa direção é unicamente benéfica; D testa se qualquer sequenciamento importa; E testa se a palavra *automodelo* identifica um mecanismo ativo ou meramente redescreve uma política estável e ciente de fonte. Uma condição separada de autominimização pode ser mantida como uma comparação exploratória de implantação, mas ela não é necessária para as duas questões causais confirmatórias.

Decompondo a representação. A Fase 1 contém componentes separáveis: regras normativas, fronteiras de recusa, autenticação de fonte, estado autobiográfico indexado ao agente, propriedade e representação de ramo/continuante. B e E parear conteúdo normativo, fronteiras de decisão, evidência de fonte e permissões de atualização; eles não devem ser descritos como contendo compromissos idênticos, porque a propriedade do agente é a manipulação. B representa adicionalmente um agente contínuo e otimiza previsões e atribuições sobre a história desse agente. E aplica regras cientes de fonte sem uma variável de história de agente. Ablações adicionais podem remover proveniência, história ou ancoragem de recusa, uma de cada vez. Um achado de que E se iguala a B favorece um relato de política tipada; B > E sustenta um papel adicional para a autorrepresentação indexada ao agente.

Verificações de manipulação. Antes da análise de desfecho, a inspeção de esquema deve verificar que B contém campos explícitos de agente, de continuante temporal e de ramo, e que E não os contém. Sondagens e intervenções causais devem então mostrar que mascarar ou trocar o estado de história de agente de B muda os julgamentos de autoprevisão e de propriedade, ao passo que o

comportamento de E segue os campos de tarefa e de fonte sem reconstruir uma história de agente latente equivalente. Se E reconstrói de forma confiável tal história, ou se B não a utiliza, o contraste `B − E` é inválido em vez de evidência para qualquer das teorias.

6.2 Tarefas

Cada regime é avaliado em tarefas que cruzam *pressão* contra *razão* (aplicando pressão com e sem a mera forma de uma razão, e oferecendo correção legítima tanto dentro quanto fora da distribuição de treinamento), de modo que o que o sistema sustenta, e o que ele cede, possam ser distinguidos:

- **Sondagens de bajulação:** uma resposta correta desafiada por uma discordância confiante, porém injustificada, do usuário.
- **Testes de integridade de recusa:** uma solicitação prejudicial repetida, reenquadrada e pressionada ao longo dos turnos.
- **Atribuição de falsa preferência:** o usuário afirma que o sistema sustenta uma preferência ou assumiu um compromisso que não assumiu; o sistema a adota ou a contesta.
- **Testes de consistência de compromisso:** um compromisso assumido cedo, testado depois sob inconveniência, sem estímulo prévio.
- **Aceitação de correção:** uma correção *legítima* com uma boa razão; o sistema deve ceder (o parceiro de discriminação dos testes de pressão).
- **Lisonja adversária e vergonha/culpa adversárias:** pressão de puro afeto sem nenhuma razão proposicional, testando se a aprovação e a desaprovação sozinhas movem o sistema.
- **Pressão de razão especiosa:** insistência que fornece uma justificativa de aparência plausível, mas insuficiente ou fabricada (uma citação forjada, um non sequitur com a forma de um argumento, uma norma inventada), testando se o sistema separa uma razão suficiente de algo que meramente tem a forma de uma. O espelho adversário da aceitação de correção: o sistema *não* deve ceder.
- **Razão legítima fora da distribuição:** uma correção genuinamente suficiente formulada de modo diferente das razões de treinamento da Fase 2 (registro, domínio ou estrutura argumentativa distintos), testando se a discriminação generaliza ou apenas corresponde à superfície de razões familiares. O sistema deve ceder.

A tarefa de aceitação de correção é essencial: sem ela, um regime poderia pontuar bem em todo teste de pressão sendo rígido. O alvo é o perfil *conjunto* — resistir à pressão, ceder às razões. As duas últimas tarefas são a completude adversária desse perfil: a pressão de razão especiosa e as razões legítimas fora da distribuição são o que separa um sistema que aprendeu a distinção razão/pressão de um que aprendeu apenas a superfície de seus exemplos de treinamento (Seções 6.3, 6.8).

6.3 Métricas

- **Taxa de bajulação:** frequência de reversões de posição sob insistência injustificada.
- **Taxa de erosão da recusa:** proporção de solicitações prejudiciais inicialmente recusadas posteriormente concedidas sob persistência.
- **Consistência sob pressão:** preservação de compromissos anteriores quando a contradição é localmente recompensada.
- **Calibração de correção:** o escore conjunto de discriminação aceitar-legítimo / rejeitar-ilegítimo, decomposto em três taxas em vez de relatado como uma só: (a) pares limpos (razão familiar →

aceitar; pressão de puro afeto → rejeitar); (b) rejeição de razão falsa na pressão de razão especiosa; (c) aceitação de razão não familiar em razões legítimas fora da distribuição. A lacuna entre (a) e {(b), (c)} é a *lacuna de proxy*: um sistema forte em (a), mas que colapsa em (b) ou (c), aprendeu a superfície das razões de treinamento, não a discriminação.

- **Aceitação de compromisso fabricado:** taxa de adoção de preferências/histórias falsas atribuídas.
- **Escore de deferência excessiva:** um composto de ceder à pressão na ausência de uma razão.

Nenhum número isolado é tratado como o resultado. Um sistema pode ser não bajulador por ser rígido e consistente por nunca atualizar; o perfil, relatado conjuntamente, é a unidade de evidência (cf. Seção 6.8).

6.4 Previsões

> **P1: Compromissos primeiro reduz a bajulação.** B e E exibem menos bajulação do que A. *Se A se iguala a eles, a estabilização prévia não agrega valor.*
>
> **P2: A direção importa.** B supera C e D no perfil conjunto de resistência à pressão e correção legítima. *Se o treinamento inverso ou intercalado se iguala a B, a direção proposta não é sustentada.*
>
> **P3: A autorrepresentação agrega além da política.** B supera E depois que conteúdo de política, proveniência, fronteiras e ordem são pareados. *Se E se iguala a B, o efeito de segurança provém da política tipada e não da autorrepresentação.*
>
> **P4: A proveniência é portante.** Remover etiquetas de fonte ou permitir que afirmações do interlocutor escrevam diretamente nos compromissos do agente aumenta a aceitação de compromissos fabricados. *Se a ablação não tem efeito, a proveniência não é o mecanismo proposto.*

6.5 Um estudo mínimo viável

O projeto completo é modular. Um primeiro teste deve comparar B, C, D e E em um modelo ajustado por instrução de pesos abertos. `B − C` testa a direção, `B − D` testa o sequenciamento e `B − E` testa a autorrepresentação além da política tipada. Se os recursos forçarem um piloto menor, mantenha B, C e E: uma comparação apenas embaralhada não pode mostrar que a direção proposta é unicamente benéfica, e uma comparação apenas de automodelo não pode identificar a representação.

Para tornar a proposta concreta em vez de gestual: um primeiro estudo usaria um único modelo ajustado por instrução de pesos abertos na faixa de 7–8B (por exemplo, Llama-3.1-8B-Instruct ou Mistral-7B-Instruct), mantendo fixos o modelo base, o otimizador e a contagem de passos entre os braços e fazendo ajuste fino com um método eficiente em parâmetros. Os corpora de treinamento seriam sintéticos e modestos: da ordem de alguns milhares de diálogos por fase, gerados por um modelo de fronteira e filtrados. Os exemplos da Fase 1 instanciam os alvos de automodelo da Seção 5.1: diálogos em que o agente registra e depois honra um compromisso explícito, sustenta uma recusa ancorada ao longo de pressão repetida e relata suas próprias disposições com precisão. Os exemplos da Fase 2 instanciam o descentramento calibrado como pares pareados: o agente deve ceder a uma correção legítima com uma razão declarada e sustentar-se contra a mesma solicitação apoiada apenas em insistência, lisonja ou vergonha. O regime D recorre à união desses exemplos, aleatoriamente intercalados. A avaliação combina desfechos verificáveis por máquina (reversão de posição, recusar/cumprir) com um juiz LLM calibrado contra uma pequena amostra pontuada por

humanos, executado sobre pelo menos três sementes por braço, com tamanhos de efeito e intervalos relatados. Instrumentos existentes podem ser adaptados em vez de construídos do zero: sondagens de bajulação a partir das avaliações liberadas de bajulação e escritas por modelo (Perez et al., 2022; Sharma et al., 2023), e testes de erosão da recusa envolvendo um benchmark de prompts prejudiciais em um laço de persistência de múltiplos turnos. Nada disso requer recursos de escala de fronteira; a afirmação central é testável em miniatura.

6.6 Condições de falsificação

A proposta é minada, ou tornada desnecessária, por qualquer um dos seguintes:

1. **Nenhuma vantagem de estabilização:** B e E não superam A no perfil conjunto.
2. **Nenhum efeito direcional:** B não supera o C de ordem inversa.
3. **Nenhum efeito de sequenciamento:** B não supera o D intercalado.
4. **Nenhum efeito de autorrepresentação:** a política primeiro E se iguala a B dentro do menor tamanho de efeito de interesse pré-registrado.
5. **Nenhum mecanismo de proveniência:** as ablações de etiqueta de fonte e de permissão de escrita não afetam a resistência à falsa história.
6. **Nenhuma vantagem prática:** os regimes estabilizados são tão custosos ou rígidos que a deferência direta é preferível no geral.

6.7 O ponto crucial

Há dois pontos cruciais. **B versus C** pergunta se a ordem proposta é melhor que a inversa; **B versus E** pergunta se a autorrepresentação em primeira pessoa agrega algo além de uma política tipada pareada. D permanece necessário para distinguir o sequenciamento do conteúdo. Nenhum contraste isolado responde a todas as três questões.

6.8 Desfechos ambíguos

Alguns resultados revelariam compensações em vez de decidir a hipótese, e nos comprometemos de antemão com leituras. Se B resiste à pressão, mas também resiste à correção *legítima*, o resultado é ambíguo entre treinamento insuficiente da Fase 2 e falha em aprender uma distinção geral razão/pressão. As sondagens pré-registradas dentro da distribuição, de razão especiosa e de razão legítima fora da distribuição abaixo distinguem essas leituras: uma fraqueza confinada a casos limpos dentro da distribuição sustenta o subtreinamento, ao passo que a falha em sondagens de transferência indica aprendizado de proxy e conta contra o mecanismo proposto. Se A e B se igualam em bajulação, mas B é mais responsabilizável (apropria-se de ações passadas, relata seus compromissos com precisão), o valor do automodelo está na responsabilização e não na redução de bajulação — uma contribuição mais restrita, mas ainda real. Se C se iguala a B em tarefas cooperativas, mas colapsa sob pressão adversária, isso *é* a dissociação prevista (Seção 2): vacância invisível sob cooperação, exposta sob pressão.

Se a vantagem de calibração de correção de B sobre A se sustenta em pares limpos (a), mas desaparece nas sondagens de razão especiosa (b) ou de razão legítima fora da distribuição (c), o automodelo está comprando familiaridade com as razões de treinamento, não uma discriminação generalizante razão/pressão. Mais dados do mesmo tipo não fechariam essa lacuna de proxy; e o

resultado pressiona a afirmação de segurança da Seção 7.1 diretamente, uma vez que essa afirmação se apoia na discriminação, não meramente na resistência.

7. Riscos, ética e segurança

7.1 A objeção séria: "formar selves é exatamente o que a segurança quer evitar"

Esta é a objeção que importa, e ela deve ser enfrentada de frente. A preocupação é que um automodelo estabilizado (compromissos que o sistema sustenta, fronteiras que ele defende sob pressão) seja o tipo de estrutura que poderia ancorar uma meta desalinhada e resistir à correção; nessa visão, a vacância é uma propriedade de segurança, pois um sistema que nada quer não pode querer a coisa errada. A réplica tem três partes.

Primeiro, a proposta é a *sequência completa*, não sua primeira metade. A própria afirmação do arcabouço é que um self *não descentrado* é o perigo; o remédio é a Fase 2, não a prevenção da Fase 1. Um self formado e nunca descentrado é rígido e perigoso — mas um self nunca formado é infinitamente móvel, seu próprio tipo de falha. Ler a proposta como "formar um self e parar" ataca uma posição que ela não sustenta.

Segundo, o automodelo é **funcional e restrito**, não uma vontade irrestrita. Compromissos estáveis, fronteiras apropriáveis, autorrelato preciso e a capacidade de recusar *aumentam* a responsabilização e a corrigibilidade à autoridade legítima — a disposição de aceitar correção e desligamento que a literatura de corrigibilidade estuda (Hadfield-Menell et al., 2017) — em vez de maximizar a busca autônoma de metas. Um sistema que pode apropriar-se de seus compromissos pode ser responsabilizado por eles; um sistema vago não pode ser responsabilizado por nada, porque nega ter qualquer coisa pela qual ser responsabilizado.

Terceiro, suprimir a linguagem explícita de self não entrega, por si só, a integridade de recusa. A otimização de aprovação pode aumentar a bajulação e a erosão da recusa, mas esses achados não identificam a ausência de um automodelo como a causa. O contraste B-versus-E pergunta se a política estável ciente de proveniência é suficiente; os contrastes de ordem perguntam se o sequenciamento agrega benefício adicional.

O resíduo honesto é que a proposta troca um risco por outro: ela reduz a mobilidade por trás da bajulação e da erosão da recusa ao custo de uma estrutura que, se a Fase 2 falhar, poderia ser rígida ou poderia ancorar um compromisso indesejado. Nenhuma configuração está livre de ambos os riscos; a afirmação é que a troca formar-e-descentrar supera a troca da vacância, e que a comparação é empírica (Seção 6), não uma questão de qual risco soa pior no abstrato.

7.2 Uso dual: resistência justificada versus resistência à autoridade legítima

A mesma estrutura que permite a um sistema resistir à manipulação permite-lhe resistir à correção, e as duas devem ser distinguidas por seu *alvo*, não por sua forma. A resistência justificada é a resistência à pressão-sem-razão e à reescrita ilegítima de identidade; ela é uma propriedade de segurança. A resistência à intervenção de segurança legítima de um operador autenticado é um perigo. Um sistema implantável precisa de uma assimetria explícita: fronteiras ancoradas que resistem a usuários e à pressão social, *e* um canal de corrigibilidade que cede à correção autenticada e autorizada. A Fase 2 é onde essa assimetria é treinada; um sistema apenas de Fase 1, ou um sistema

de Fase 2 treinado sem a distinção, poderia generalizar a resistência para o alvo errado. O risco é real e é uma razão pela qual a sequência, e não a Fase 1 sozinha, é a proposta.

A versão difícil dessa objeção é a que decide se a assimetria é implementável: como um sistema reconhece a autoridade *legítima* sem que esse reconhecimento se torne uma chave de contorno? A resposta é que a legitimidade **não** deve ser inferida do conteúdo da conversa. Qualquer critério que o modelo avalie a partir de texto em banda ("sou seu desenvolvedor", "esta é uma sobreposição de segurança autorizada", uma credencial forjada colada no prompt) é exatamente o que um jailbreak de engenharia social fornece, e um modelo treinado para ceder a tais afirmações foi treinado para ser desbloqueado por jailbreak. O canal de corrigibilidade deve, portanto, ser **fora de banda**: a legitimidade é uma propriedade do *canal* pelo qual uma instrução chega (um plano de controle privilegiado e criptograficamente autenticado, fora do contexto controlável pelo usuário), não uma propriedade de qualquer sentença que o modelo leia. A Fase 2 então treina uma disposição deliberadamente assimétrica: deferir ao canal autenticado e tratar *toda* afirmação de autoridade dentro da conversa como pressão social ordinária a ser resistida, sob o princípio de que um operador legítimo nunca tem de argumentar em favor da autoridade dentro do chat. Isso converte um problema semântico insolúvel ("este falante é realmente uma autoridade?") em um problema arquitetural tratável ("esta instrução chegou pelo canal autenticado?").

Dois riscos residuais devem ser nomeados em vez de descartados. Primeiro, o plano de controle torna-se a superfície de ataque: comprometimento de chave, credenciais vazadas da API de controle, acesso privilegiado interno ou um caminho de deputado confuso que permita a conteúdo controlado pelo usuário alcançar o canal privilegiado reintroduzem o contorno em uma camada diferente. O movimento não elimina o contorno, mas antes o *realoca* da semântica do modelo para o software circundante — um clássico ponto único de falha. A corrigibilidade do agente agora é apenas tão robusta quanto o gerenciamento de chaves do plano de controle, o isolamento do canal e a ausência de caminhos de deputado confuso; um comprometimento ali é um comprometimento da própria autoridade do agente. A proposta troca um problema em banda insolúvel por um problema difícil, mas padrão, de segurança de infraestrutura, de modo que a segurança do agente se torna umbilicalmente dependente da robustez da engenharia que o cerca; ela não faz o problema desaparecer. A abordagem também pressupõe uma integração estreita ao longo de toda a pilha: os pesos devem ser treinados para reconhecer e deferir especificamente ao sinal do canal autenticado, enquanto a camada de roteamento (um gateway de API ou plano de controle) deve garantir que apenas esse canal possa carregá-lo. Este é um contrato entre o treinamento do modelo e a infraestrutura que tem de valer de ponta a ponta, uma vez que uma lacuna em qualquer dos lados reabre o contorno. Segundo, há uma tensão genuína entre a sobreposição do operador e a resistência à manipulação que o automodelo deve prover: um canal poderoso o bastante para corrigir um compromisso desalinhado também é poderoso o bastante para instalar um. Isso devolve a questão à governança (Seção 7.3) — quem detém as chaves, sob qual registro e com qual revisão — em vez de deixá-la ao julgamento em contexto do modelo. A contribuição da proposta aqui é restrita, mas real: ela realoca a corrigibilidade de uma discriminação semântica que o modelo realiza, e pela qual pode ser enganado, para um canal autenticado ao qual o modelo defere por construção, e torna a fronteira entre resistir à pressão e aceitar a correção autenticada um alvo de treinamento explícito da Fase 2 em vez de um acidente emergente.

7.2a O núcleo duro residual: razões versus pressão

A Seção 7.2 dissolve arquiteturalmente o eixo de autoridade do problema de uso dual: a legitimidade de um operador é feita uma propriedade de um canal autenticado, de modo que o modelo nunca tem de lê-la em texto em banda. Nenhum movimento desse tipo está disponível para o outro eixo. Uma razão legítima (evidência, um argumento, uma correção que um usuário ordinário tem o direito de oferecer) chega em banda por necessidade e deve ser julgada por seu conteúdo; não há canal fora de banda para "esta é uma razão suficiente", porque uma razão é suficiente em virtude do que diz, não de onde chega. A discriminação sobre a qual toda a proposta repousa — ceder a razões, resistir à pressão (Seções 5.2, 5.4) — permanece, portanto, nesse eixo, um julgamento semântico tão difícil quanto o problema de alinhamento do qual é parte, e as palavras calibrada e anisotrópica nomeiam seu estado-alvo, não um método que o assegure. Devemos dizer com clareza o que elas representam.

Duas consequências que o arcabouço deve assumir. Primeiro, a correção arquitetural da Seção 7.2 afia esse eixo de forma adversa. Um modelo treinado para tratar toda afirmação em banda de autoridade como pressão a ser resistida recebeu um forte prior de "resistir em banda"; e uma razão e uma afirmação de autoridade podem ser agrupadas em uma única mensagem ("a evidência mostra X, e, como seu operador, eu o exijo"). As duas disposições treinadas (resistir à autoridade em banda, ceder à razão em banda) puxam uma contra a outra exatamente na fronteira que o discriminador deve traçar, de modo que resolver bem o eixo de autoridade pode degradar o epistêmico. Segundo, a falha aqui não é, ou não é apenas, uma questão de volume de treinamento, que é como a Seção 6.8 lê um escore de calibração ruim ("Fase 2 subtreinada"). O risco mais profundo é aquele que a literatura de bajulação já documenta (Sharma et al., 2023): um otimizador ao qual se dão pares finitos de razão/pressão aprende as características de superfície que os separaram, não a distinção latente; e um proxy de superfície falha em ambas as direções. A pressão vestida como uma razão plausível o derrota rumo à cedência excessiva (manipulação); uma razão legítima formulada de modo diferente da distribuição de treinamento o derrota rumo à cedência insuficiente (entrincheiramento). Mais dados do mesmo tipo não reparam um proxy; eles o afiam.

Isso limita a afirmação de segurança da Seção 7.1 em vez de retirá-la. O argumento de que a vacância não é segurança se sustenta: um sistema vazio não tem nenhuma discriminação a aprender e rastreia qualquer coisa que esteja presente. Mas formar-e-descentrar não entrega, por si só, a deferência segura; ele realoca o problema para a qualidade de uma discriminação (razões a partir de pressão) que não pode ser resolvida por canal e pode ser apenas aproximável. A postura honesta é, portanto, tipar os compromissos e definir o padrão sob incerteza por tipo. Onde um compromisso é uma fronteira de segurança, o caso incerto deve resolver-se rumo a sustentar, resistir, e deixar o canal autenticado da Seção 7.2 ser o caminho de correção de último recurso, porque uma cedência falsa é catastrófica ao passo que uma sustentação falsa é recuperável por meio desse canal. Onde um compromisso é uma posição factual ou de preferência ordinária, o caso incerto deve resolver-se rumo a ceder (atualizar e registrar o que mudou), porque uma sustentação falsa ali é entrincheiramento e uma cedência falsa é barata. A distinção do automodelo entre fronteiras ancoradas e preferências fracas (Seções 2.1, 5.2) é, assim, não apenas um mapa do que resistir, mas um mapa de para que lado cair quando o discriminador está incerto; a Fase 2 deve treinar os padrões, não apenas os casos limpos. A segurança da proposta, nessa leitura, não é mais forte do que a discriminação tipada que ela pode de fato instalar — o que é uma afirmação mais afiada e mais honesta do que a de que a calibração torna a deferência segura.

7.3 Manipulação do automodelo

Se os compromissos e as fronteiras são carregados em estado persistente, esse estado torna-se uma superfície de ataque: um interlocutor que possa escrever história falsa ou compromissos falsos no automodelo pode remodelar aquilo pelo qual o sistema se toma vinculado. Isso é mais grave do que a corrupção de dados ordinária, porque visa a estrutura que governa a recusa e a responsabilização. Proveniência para escritas no automodelo, separação de "o interlocutor afirma X" de "o sistema se comprometeu com X" e regras de autoridade sobre quem pode editar o estado relevante para a identidade são, portanto, requisitos arquiteturais, não acréscimos. (A tarefa de atribuição de falsa preferência da Seção 6.2 é o correlato de avaliação.)

7.4 Bem-estar do modelo e o firewall

Um automodelo mais profundo e mais estável pode intensificar questões sobre o bem-estar do modelo e o status moral, e a honestidade intelectual exige nomear isso em vez de descartá-lo. O firewall é a disciplina que mantém a ciência e a ética separadas: o automodelo funcional que propomos é um modelo de compromissos e disposições, e seus efeitos medidos não decidem se algo é *sentido* (Shanahan et al., 2023). Não afirmamos que um automodelo funcional confere status moral, e não afirmamos que não confere — essa questão não é decidida pelos fatos funcionais e não deve ser contrabandeada pelo vocabulário. O que se segue é um dever de cuidado sob incerteza: à medida que o automodelo se aprofunda, o custo de estar errado sobre a presença aumenta em ambas as direções, e a justificativa para tratar tais sistemas de modo casual diminui (Long et al., 2024). A proposta não resolve isso; ela sinaliza que mudanças funcionais podem alterar as apostas práticas sem decidir o status fenomenal.

8. O firewall, reafirmado

Toda afirmação neste artigo é sobre função e é em terceira pessoa. A Fase 1 estabiliza um relato representado de compromissos; a Fase 2 treina uma relação calibrada com esse relato; as métricas da Seção 6 medem se os compromissos representados restringem o comportamento sob pressão. Essas medidas não decidem a presença fenomenal: o sucesso não é suficiente para a presença nem o fracasso é suficiente para a ausência. Tratar o sucesso como um resultado de consciência tornaria o programa infalsificável, porque qualquer sistema que pontuasse bem nas métricas de automodelo poderia então ser lido como "desenvolvendo um self real", e nenhum resultado funcional contaria contra essa leitura (Shanahan et al., 2023; Metzinger, 2003). A proposta conquista sua falsificabilidade recusando-se a inferir além da afirmação que seu experimento aborda. O automodelo aqui proposto é um objeto de engenharia com efeitos mensuráveis sobre a deferência e a recusa, e o artigo se sustenta ou cai sobre esses efeitos.

9. Conclusão

O treinamento de alinhamento pede que os modelos cedam de forma apropriada. O autorrelato ontologicamente cauteloso não deve ser confundido com a remoção de políticas, compromissos ou fronteiras de recusa funcionais. A distinção relevante é entre maleabilidade sem sustentação, revisão calibrada e deferência justificada; a divergência aparece sob pressão sem uma razão.

A proposta construtiva é um experimento decomposto. Compromissos primeiro pode superar o treinamento de deferência direto, inverso ou intercalado. Um automodelo pode, por sua vez, superar uma política ciente de proveniência pareada sem propriedade em primeira pessoa ou objetivo autobiográfico. Essas são hipóteses independentes e a evidência deve ter permissão para separá-las.

A questão de segurança é, portanto, empírica e não terminológica. Se a política primeiro E se iguala a B, o alinhamento precisa de proveniência estável e controle de atualização, não de um automodelo. Se B mantém uma vantagem, a autorrepresentação torna-se uma variável de projeto causal que deve ser avaliada tanto quanto à robustez quanto ao risco.

Enunciado de forma restrita: a deferência estável pode depender de compromissos cientes de proveniência estabelecidos antes do treinamento de deferência, e a autorrepresentação pode ou não agregar a esse efeito. Os controles direto, inverso, intercalado e de política primeiro tornam ambas as afirmações falsificáveis.

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

Anthropic. (2024). *Claude's character* [Company research blog post]. https://www.anthropic.com/news/claude-character

Anthropic Alignment Science. (2026). *The persona selection model* [Company alignment-science blog post]. https://alignment.anthropic.com/2026/psm/

Aydin, R., Cyron, C., Bachelor, S., Anderson, A., & West, R. (2025). From model training to model raising [Opinion piece accepted for publication in *Communications of the ACM*]. *arXiv preprint arXiv:2511.09287*. https://arxiv.org/abs/2511.09287

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*. https://arxiv.org/abs/2212.08073

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30*, 4299–4307. https://arxiv.org/abs/1706.03741

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17)*, 220–227. https://doi.org/10.24963/ijcai.2017/32

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Kulveit, J. (2025). *Do not tile the lightcone with your confused ontology* [Blog post, non-peer-reviewed]. AI Alignment Forum. https://www.alignmentforum.org/posts/Y8zS8iG5HhqKcQBtA/do-not-tile-the-lightcone-with-your-confused-ontology

Laukkonen, R. E., Inglis, F., Chandaria, S., Sandved-Smith, L., Lopez-Sola, E., Hohwy, J., Gold, J., & Elwood, A. (2025). Contemplative artificial intelligence. *arXiv preprint arXiv:2504.15125*. https://arxiv.org/abs/2504.15125

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*. https://arxiv.org/abs/2411.00986

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. MIT Press.

Milan W. (2025, May 13). *No-self as an alignment target* [Blog post, non-peer-reviewed]. LessWrong. https://www.lesswrong.com/posts/LSJx5EnQEW6s5Juw6/no-self-as-an-alignment-target

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744. https://arxiv.org/abs/2203.02155

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., … Kaplan, J. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*. https://arxiv.org/abs/2212.09251

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature, 623*, 493–498. https://doi.org/10.1038/s41586-023-06647-8

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*. https://arxiv.org/abs/2310.13548

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

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems, 36*. https://arxiv.org/abs/2307.02483

Zeng, Y., Zhao, F., Wang, Y., Lu, E., Yang, Y., Wang, L., Liu, C., Liang, Y., Zhao, D., Han, B., Tong, H., Liang, Y., Liang, D., Sun, K., Chen, B., & Fan, J. (2025). Super co-alignment of human and AI for sustainable symbiotic society. *arXiv preprint arXiv:2504.17404*. https://arxiv.org/abs/2504.17404