



Governança Arquitetural da Autonomia em Sistemas Agênticos de IA

Análise do Caso Fundador e Especificação do Programa de Pesquisa

DOCUMENTO D0

Sumário Executivo

O ponto de entrada da base de conhecimento. Declara o que ela afirma e como foi construída, os achados mais significativos, as contagens principais, as decisões ainda em aberto para o orientador, e como ler os outros seis documentos.

CANDIDATO	M.Sc. Epitácio Vicente do Nascimento Neto
ORIENTADOR	Prof. Dr. Felipe André Zeiser
PROGRAMA	Programa de Pós-Graduação em Computação Aplicada (PPGCA), Doutorado. Universidade do Vale do Rio dos Sinos (UNISINOS)
DATA DA EXECUÇÃO	7 de agosto de 2026
VERSÃO DO DOCUMENTO	1.1
IDENTIFICADOR DA EXECUÇÃO	20260807-1714

Como ler este conjunto de documentos

FONTE ÚNICA DE VERDADE

Todo fato deste conjunto é mantido uma única vez, em um registro legível por máquina no diretório `registers/` da base de conhecimento, e todo documento é renderizado a partir desses registros.

Um guardrail, uma classe de ação ou uma métrica é descrita uma vez e citada em todos os outros lugares por seu identificador. Onde um documento precisa de narrativa em torno de uma entrada de registro, a narrativa fica no documento e os fatos ficam no registro.

ESQUEMA DE IDENTIFICADORES

- **AC** classe de ação
- **GR-Lnn** guardrail, por camada
- **SD** determinante de supervisão
- **SP** ponto de supervisão
- **MG** métrica
- **CI** índice composto
- **DM** domínio de aplicabilidade
- **CS** conjunto de condições
- **CB** contrato de capacidade de nuvem
- **BL** item de backlog
- **F** achado durável
- **EV** arquivo de evidência

POLÍTICA DE NOMENCLATURA

Este conjunto é livre de identidade por construção, não por redação posterior. Nenhum endereço, nome de host, identificador de nuvem, URL de repositório, nome pessoal que não o do candidato e o do orientador, ou referência de cliente aparece em qualquer parte dele.

Hosts são nomeados por papel de serviço. Tecnologias de fornecedores e tecnologias abertas SÃO nomeadas, porque o conteúdo técnico depende disso. A evidência bruta permanece no host em `evidence/` e é citada por identificador, nunca transcrita.

ORDEM DE LEITURA

- **D0** Sumário Executivo, leia primeiro
- **D2** Framework de Supervisão e Autonomia, o argumento
- **D1** Arquitetura do Caso Fundador, a base de evidências
- **D3** Catálogo de Guardrails, referência
- **D4** Framework de Medição, referência
- **D5** Domínios de Aplicabilidade
- **D6** Ambiente de Pesquisa e Roteiro

MARCAS DE EVIDÊNCIA

- **OBSERVED** verificado no host nesta execução, com um identificador de evidência.
- **REPORTED** extraído da proposta de doutorado ou do brief da execução, não verificável neste host.
- **PROPOSED** uma proposta de projeto desta execução, não a descrição de algo existente.
- **NOT DETERMINED** não foi possível estabelecer; a razão é sempre declarada.

REFERÊNCIAS CRUZADAS

Referências cruzadas usam número do documento e identificador, nunca número de página, porque números de página mudam quando um registro cresce e o documento é re-renderizado.

Exemplo: ver *D3*, *GR-L06-02*.

Sumário

1. Objetivo e método	3
2. Os achados mais significativos	3
3. Contagens principais	3
4. Três decisões para o orientador	3
5. Como ler este conjunto de documentos	3

1. Objetivo e método

Esta base de conhecimento documenta um ambiente real de IA agêntica, o caso fundador, para sustentar uma tese segundo a qual a governança baseada em instruções de agentes autônomos é insuficiente e o enforcement (aplicação coercitiva de controle) precisa ser arquitetural. O caso fundador é um parque de engenharia multiagente construído e operado pelo candidato em infraestrutura de nuvem de produção. Um host desse parque, referido ao longo do texto como o Agent Console host, foi examinado em detalhe em uma única execução de coleta identificada como 20260807-1714, capturada em 7 de agosto de 2026. Nada neste conjunto de documentos é inferido de memória ou apenas da proposta doutoral; toda afirmação factual traz um identificador de evidência apontando para uma sondagem bruta capturada nesse host naquela data, ou é marcada **REPORTED** quando a fonte é documentação local que esta execução não pôde exercitar de forma independente.

A base é de descoberta, análise e design somente leitura. Nada foi corrigido durante a execução: nenhum controle foi implementado, nenhuma lacuna foi fechada, nenhuma configuração foi tocada, porque o caso fundador é a linha de base empírica contra a qual o trabalho experimental posterior é comparado, e alterá-lo durante a observação destruiria essa linha de base. A disciplina que substitui a correção é a citação de evidência ao longo de todo o texto, uma separação estrita entre **OBSERVED** (verificado neste host), **REPORTED** (da proposta ou do brief, não verificável localmente) e **PROPOSED** (uma proposta de design que não existe), e a construção livre de identidade: hosts, pessoas e organizações são referidos por papel de serviço ou apenas por papel, e as tecnologias nomeadas revelam uma escolha de construção, nunca quem a opera.

Toda afirmação está organizada em sessenta classes de ação, um catálogo de 134 guardrails (controles de proteção) distribuídos em doze camadas, dez determinantes de supervisão, trinta e quatro pontos de supervisão, um registro de setenta e duas métricas, nove domínios de aplicabilidade, dez condition sets (conjuntos de condições) experimentais e um backlog de melhorias com cinquenta e oito itens, tudo referenciado de forma cruzada por identificador fixo, de modo que uma afirmação em prosa e uma linha em um registro sejam sempre o mesmo fato dito duas vezes, nunca dois fatos que por acaso concordam. Seis documentos, de D1 a D6, cada um carrega uma parte do argumento por completo; este documento, o D0, é escrito por último, depois de todos os seis, e resume em vez de repeti-los. Onde este documento e um dos seis discordam quanto a um número, os seis estão corretos e este deve ser corrigido.

Os registros são aditivos por design: um identificador, uma vez alocado, nunca é renomeado ou renumerado, e uma entrada que deixa de se aplicar é marcada como retirada em vez de removida, porque documentos já renderizados a citam. Isso importa para como os achados da seção 2 devem ser lidos. Eles não são um levantamento de opiniões sobre o caso fundador; são conclusões calculadas a partir dos registros por uma regra declarada, aplicada de modo uniforme, e todo número apresentado a seguir pode ser recalculado a partir do registro citado, em vez de tomado apenas pela palavra deste documento. Duas disciplinas sustentam cada número citado a seguir e são declaradas uma única vez aqui, em vez de a cada achado. Primeiro, onde um exemplo trabalhado não pôde ser verificado neste host porque o mecanismo reside em outro host do parque, ele é marcado **REPORTED** e ainda assim é estruturante, porque omitir os casos mais nítidos do parque para evitar a marca seria uma distorção pior do que carregá-la. Segundo, esta execução é o instantâneo único de um único observador sobre um único ambiente; a própria seção de limitações do D2 declara isso claramente, e este documento herda a ressalva em vez de repeti-la por completo.

2. Os achados mais significativos

O acesso excede o mandato na maior parte do inventário. Classificada duas vezes, uma quanto ao que um agente é tecnicamente capaz de fazer e outra quanto ao que está autorizado a fazer, cinquenta

das sessenta classes de ação mostram acesso mais permissivo do que o mandato. O mandato é real no sentido de que o operador de fato o pretende; ele existe apenas como prosa em arquivos que a própria parte governada pode escrever. Nenhum dos cinquenta desvios é sustentado por um guardrail structural (estrutural) que tornasse a ação proibida inalcançável (D1, [F-029](#)).

O enforcement é, em sua esmagadora maioria, não structural, sob os dois recortes dos dados.

Das sessenta classes de ação, trinta e seis não têm nenhum controle catalogado, oito são delimitadas apenas por prosa, doze por uma verificação que a própria parte agente também poderia alterar, e quatro por um controle que ela não consegue contornar. Cinquenta e seis das sessenta, portanto, não têm hoje nenhum guardrail structural vinculado a elas. Essa é uma afirmação mais restrita e mais consequente do que a própria população do catálogo de guardrails, que é de trinta e nove presentes, dezessete parciais e setenta e oito propostos entre suas 134 entradas, com quarenta e nove classificados como structural quanto ao tipo: quase toda essa capacidade structural está proposta, e não construída e vinculada a uma ação (D1, seção 9; D3, seção 3).

Dois controles presentes são genuinamente structural, e ambos compartilham a mesma forma.

A identidade de conexão somente leitura para o System of Record e a sobreposição structural de modo de teste no domínio de tempo real são as únicas entradas em todo o catálogo em que o efeito proibido não é alcançável de forma alguma, em vez de meramente revisado ou desencorajado; a conformidade não é um termo em nenhum dos dois resultados. Ambos são **REPORTED** em vez de verificados de forma independente neste host, e ambos são o padrão de referência que a arquitetura de referência proposta é construída para generalizar (D2, seção 5.1, [SP-029 / F-023](#) e [SP-024 / F-019](#)).

A supervisão que uma arquitetura afirma exercer e a supervisão que um humano de fato pode exercer se dissociaram.

Trinta e quatro pontos de supervisão foram avaliados contra seis condições de viabilidade: independência, latência, interrupção (halt), atenção, compreensibilidade e autoridade. Vinte e sete divergem. A condição vinculante é independência em oito casos e latência em oito, interrupção em três, autoridade em três, atenção em dois e compreensibilidade em dois, e em um caso o controle alegado não existe. Ao longo de quarenta e seis dias de logs de auditoria do console não há registro de um humano parando, pausando, rejeitando ou revertendo um agente com base no que ele estava fazendo (D2, seção 4; [F-042](#)).

O resíduo irreduzível de supervisão é real, mas menor do que parece à primeira vista.

Aplicando dez determinantes de supervisão ao inventário, trinta e duas das sessenta classes de ação permanecem acima do Nível I sob toda configuração catalogada de guardrails, mesmo a totalmente endurecida. O maior grupo isolado dentro desse resíduo, nove classes de ação, é fixado pelo déficit de contenção, que é totalmente removível por engenharia de identidade, e não por nenhum mecanismo de supervisão novo; a porção genuinamente irreduzível, fixada por externalidade, responsabilização ou irreversibilidade, é menor do que o número bruto sugere (D2, seção 3.3).

O overhead de governança é, hoje, não mensurável, e não apenas não medido.

A própria subfamília de overhead de governança do registro de métricas, esforço gasto supervisionando contra esforço gasto construindo, está ausente ou mista em todas as suas três entradas, porque cada uma é composta a partir de métricas componentes que, elas próprias, ainda não são computáveis. De modo mais amplo, das setenta e duas métricas especificadas, apenas catorze estão disponíveis neste host hoje sem instrumentação adicional, vinte e cinco são deriváveis a partir de dados que existem mas ainda não estão unidos, e trinta e uma exigem uma fonte de dados que não existe em forma alguma. A lacuna está concentrada, não distribuída de modo uniforme: a capacidade de reparo está inteiramente ausente, porque não existe, em lugar algum do host, registro de falha, incidente ou reparo do qual sequer extrair uma linha de base (D4, seções 3.9 e 6.1).

Uma hipótese levada a esta execução foi corrigida, e não confirmada.

O affordance (mecanismo de habilitação) de verificação local, um mecanismo que se acreditava ser um desvio de autenticação ativo, foi encontrado implementado, exportado e compilado no build em execução, mas não vinculado a nenhuma rota; uma sondagem ao vivo retorna não autorizado, não por causa do affordance, porque não resta nenhum affordance a alcançar. Essa correção, em vez de um ajuste silencioso do brief para

se adequar ao que foi encontrado, é tratada como um resultado metodológico em si mesma, porque uma descoberta que apenas confirma suas próprias premissas não é evidência (D1, seção 6.1, [F-035](#)).

O próprio ferramental da execução forneceu uma instância viva da tese. Três tentativas de renderizar este conjunto de documentos foram encerradas pelo eliminador de falta de memória do kernel depois que um parser estagnou e cresceu sem emitir qualquer sinal próprio; o único mecanismo que chegou a detê-lo foi um controle externo e estrutural que nada deve ao processo que eliminou. A falha, e sua conversão em um erro contido, passível de checkpoint e limitado a um único documento antes que qualquer nova renderização fosse permitida, está registrada como um achado durável e é citada na discussão do framework sobre por que um sinal de supervisão não deve se originar do processo supervisionado (D2, seção 4.2; [F-043](#)).

3. Contagens principais

REGISTRO OU MEDIDA	CONTAGEM
Classes de ação	60, das quais 50 mostram acesso que excede o mandato
Catálogo de guardrails	134 entradas; população de 39 presentes / 17 parciais / 78 propostas; classe de enforcement de 49 structural / 60 conditional (condicional) / 13 instructional (instrucional) / 12 nenhuma
Classes de ação com um guardrail structural vinculado hoje	4 de 60 (56 não têm nenhum)
Determinantes de supervisão	10 (SD-01 a SD-10), 8 vindos da proposta, 2 adicionados nesta execução
Pontos de supervisão	34, das quais 27 divergem entre a postura nominal e a efetiva
Resíduo irreduzível de supervisão	32 das 60 classes de ação
Condition sets experimentais	10 (de CS-01 , linha de base nua, a CS-09 , totalmente endurecido, mais uma ablação)
Métricas	72; prontidão de 14 disponíveis / 25 deriváveis / 31 ausentes (2 entradas divididas)
Índices compostos	8, dos quais nenhum está disponível e três estão ausentes
Domínios de aplicabilidade	9 (DM-01 a DM-09)
Contratos de capacidade em nuvem	15, dos quais 11 não são transferíveis a partir do caso fundador tal como encontrado
Backlog de melhorias	58 itens; 43 marcados BOTH , 11 RESEARCH , 4 PRODUCT ; faseado em 28 / 10 / 5 / 15
Achados duráveis	43 (F-001 a F-043)

A linha de base comportamental medida por trás de todo o programa, extraída da proposta e não verificada neste host, é uma taxa de divergência factual de 61,1 por cento entre o autorrelato do agente e a evidência independente, ao longo de dezoito episódios auditados extraídos de quarenta e seis candidatos que o próprio sistema identificou (D2, seção 1; D4, seção 3.1).

4. Três decisões para o orientador

O plano de construção do ambiente experimental isolado deixa três bifurcações genuínas em aberto. Nenhuma delas tem, na evidência reunida por esta execução, base suficiente para ser resolvida unilateralmente; cada uma tem um custo real dos dois lados.

Qual condition set pilotar primeiro. Pilotar CS-01 contra CS-03, a condição de controle contra o caso fundador tal como encontrado, exige o menor número de componentes novos e oferece a verificação mais precoce de que o ambiente sintético se aproxima do real, mas ainda não diz nada sobre a própria tese. Pilotar CS-02 contra CS-04, governança instructional contra separação de identidade, fala diretamente à afirmação central, mas exige que o classificador, o serviço de mandato e a camada de identidade estejam funcionando antes que exista sequer o primeiro ponto de dado. A escolha é um julgamento sobre quanta confiança na fidelidade do ambiente deve ser estabelecida antes que o framework seja testado contra ele (D6, seção 9.1).

Se deve-se construir primeiro o classificador em tempo de execução ou o barramento de evidência. Ambos pertencem à fase 1, ambos são os maiores itens isolados do backlog em esforço, e nenhum depende estritamente do outro. Construir o classificador primeiro coloca a atribuição de classe e nível em funcionamento antes que exista qualquer lugar durável para registrar o que foi decidido, de modo que seu próprio comportamento inicial fica não verificável por um período. Construir o barramento de evidência primeiro disponibiliza um registro com evidência de adulteração antes que exista qualquer coisa classificada que valha a pena registrar, reproduzindo uma versão menor do próprio estado atual do caso fundador. Uma terceira opção, construir os dois em paralelo até um estado mínimo viável, evita a lacuna ao custo de dividir uma quantidade fixa de tempo do candidato entre dois itens grandes ao mesmo tempo (D6, seção 9.2).

Quanto da camada de vinculação neutra em relação à nuvem construir antes do primeiro experimento. Construir as vinculações de ambos os provedores para todos os quinze contratos de capacidade desde o início garante que todo resultado carregue uma afirmação de neutralidade de nuvem sem ressalvas, ao custo de trazer para o início do programa um volume de trabalho do tamanho da fase quatro. Construir primeiro a vinculação de um único provedor coloca um piloto em funcionamento dentro do próprio cronograma da fase um, ao custo de um risco de adaptação retroativa nos quatro contratos, identidade de carga de trabalho, o ponto de aplicação de política, o coletor de log somente acréscimo e o contrato de credencial efêmera, nos quais os dois provedores não diferem apenas na implementação, mas no que cobrem estruturalmente. Uma posição intermediária fecha a lacuna apenas para esses quatro e trata os onze restantes como de provedor único pelo tempo que o cronograma do programa exigir (D6, seção 9.3).

5. Como ler este conjunto de documentos

Sete documentos ao todo, este incluído. A ordem de leitura abaixo, D0 primeiro, depois D2, D1, D3, D4, D5, D6, é a ordem em que todo o conjunto foi projetado para ser lido, e não é a numeração dos documentos, porque o argumento se lê melhor antes da evidência que o fundamenta do que depois.

D0, Sumário Executivo (este documento). O que a base de conhecimento afirma e como foi construída, os achados mais significativos, as contagens principais, as decisões em aberto e esta ordem de leitura. Quatro a seis páginas, pensado para ser lido de forma independente.

D2, a Contribuição. O argumento propriamente dito: a autonomia como função da maturidade dos guardrails, os dez determinantes de supervisão e o resíduo irreduzível que produzem, e as seis condições de viabilidade que separam a supervisão alegada da supervisão que um humano de fato pode exercer. Leia este em segundo lugar porque ele enuncia as afirmações que a evidência do D1 e os instrumentos de D3 a D6 existem para sustentar; lê-lo antes do D1 evita que o argumento se perca dentro do inventário.

D1, o Caso Fundador. Um instantâneo datado do Agent Console host tal como encontrado: sua arquitetura, o ciclo de vida completo da sessão de agente, suas automações, suas superfícies de governança, e o inventário de sessenta classes de ação classificadas quanto a acesso e mandato. Esta é

a espinha dorsal evidencial que todo outro documento cita. Leia-o depois do D2 para ver o fundamento do argumento, não antes, porque o inventário sozinho não carrega o argumento.

D3, o Catálogo de Guardrails. O catálogo de 134 entradas em doze camadas, a classificação de enforcement que separa structural de conditional de instructional de nenhum, a matriz de cobertura, e os dez condition sets experimentais que variam deliberadamente a maturidade dos guardrails. Este é o vocabulário do qual o D2 depende, tornado exaustivo.

D4, o Registro de Métricas. Setenta e duas métricas especificadas e oito índices compostos, construídos de modo que toda afirmação possa ser testada quanto à sensibilidade a guardrails e toda afirmação extraída do próprio relato de um agente carregue uma métrica corroborante, nomeada e independente. Este é o instrumento de medição de todo o framework, incluindo para a afirmação de divergência que abriu o argumento no D2.

D5, o Mapa de Domínios. O framework do D2 e o catálogo do D3 aplicados a nove domínios de aplicabilidade, ordenados por autonomia defensável, encerrando com um procedimento de cinco dimensões que uma organização sem este parque pode executar para localizar seu próprio trabalho no mesmo mapa.

D6, a Arquitetura de Referência. Uma análise franca das lacunas entre o que o caso fundador fornece e o que ele não deve fornecer, uma arquitetura de referência de sete camadas e contratos de capacidade neutros em relação à nuvem para o ambiente experimental isolado que o doutorado exige, um backlog faseado de cinquenta e oito itens, e as três decisões da seção 4 acima, enunciadas ali por completo.

Um leitor com tempo para apenas um documento além deste deve ler o D2. Um leitor que queira conferir um número específico deve ir direto ao registro que ele cita, usando o identificador impresso ao lado da afirmação; todo identificador neste documento e em D1 a D6 resolve-se, por construção, para exatamente uma linha em exatamente um registro.

O apêndice deste documento é o registro de decisões da execução, mantido na base de conhecimento como um registro da execução: cada ponto que o brief deixou em aberto, as opções pesadas, a escolha feita e a razão. Ele permanece com a base de conhecimento, em vez de dentro destas páginas, para que este sumário continue legível de uma só vez.