UNISINOS

Governança Arquitetural da Autonomia em Sistemas Agênticos de IA

Análise do Caso Fundador e Especificação do Programa de Pesquisa

DOCUMENTO D3

Catálogo de Guardrails

> *O catálogo de guardrails (controles de proteção). Define doze camadas arquiteturais, classifica cada controle catalogado conforme restrinja um agente de modo structural (estrutural), conditional (condicional) ou apenas instructional (instrucional), e deriva os dez condition sets (conjuntos de condições) experimentais usados para variar a maturidade dos guardrails de forma controlada.*

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. Como ler o catálogo

Este documento é um registro com narrativa ao redor, não um argumento com evidência anexada. A afirmação que ele existe para sustentar, de que a governança baseada em instrução é insuficiente e o enforcement (aplicação coercitiva de controle) precisa ser arquitetural, é feita no D2. O que este documento fornece é o vocabulário do qual o D2 depende: um conjunto fixo de camadas, uma classificação fixa do que um controle de fato faz quando um agente tenta agir contra ele, e um conjunto fixo de condition sets experimentais para que a maturidade dos guardrails possa ser variada deliberadamente, em vez de descrita de forma impressionista.

O catálogo contém 134 entradas, identificadores `GR-L01-01` a `GR-L12-09`, distribuídas em doze camadas. Trinta e nove estão presentes hoje no Agent Console host, dezessete são parciais, setenta e oito são propostas. Quarenta e nove são classificadas structural, sessenta conditional, treze instructional e doze não emitem controle algum sob a classificação definida na Seção 1.2. Esses dois últimos fatos não descrevem os mesmos quarenta e nove controles: das trinta e nove entradas de fato presentes no host, apenas duas são structural. O restante da Seção 1 explica por que essa lacuna entre "existir" e "aplicar de modo structural" é o fato organizador de todo este catálogo.

1.1 O modelo de camadas

Doze camadas, `L1` a `L12`, ordenadas desde a identidade sob a qual um processo é executado até a prosa que um agente lê como instrução. A ordenação não é arbitrária: ela vai da camada mais distante do modelo, onde um controle não depende de o agente compreendê-lo ou concordar com ele, até a camada mais próxima do modelo, onde um controle não é nada além de compreensão e concordância.

- **L1 Identidade e credencial.** Quem ou o que o agente é, criptograficamente, e o que essa identidade pode obter.
- **L2 Rede e saída.** O que os processos do agente conseguem alcançar em uma rede, tanto de entrada quanto de saída.
- **L3 Sistema de arquivos e espaço de trabalho.** O que o agente pode ler, escrever ou isolar em disco.
- **L4 Processo e execução.** Como os processos do agente são criados, limitados e encerrados.
- **L5 Plano de dados.** O que o estado armazenado do agente, e os sistemas de registro que ele toca, podem representar ou permitir de modo structural.
- **L6 Mudança e implantação.** Como uma mudança chega a um sistema em execução, e quem ou o que a controla.
- **L7 Mandato e política.** O que o agente está autorizado a fazer, como um artefato verificável por máquina, e não como prosa.
- **L8 Evidência, auditoria e verificação.** Que registro existe do que aconteceu, e quem pode verificá-lo de forma independente do agente que atuou.
- **L9 Contenção e interrupção.** O que interrompe um agente, parcial ou completamente, depois que ele já está agindo.
- **L10 Interface humana e supervisão.** O que um humano de fato vê e pode fazer a respeito.
- **L11 Temporal e contextual.** Que restrições de tempo, sequência e duração se aplicam a uma ação.
- **L12 Camada de modelo e prompt.** O que é dito ao agente, em linguagem que ele lê como contexto.

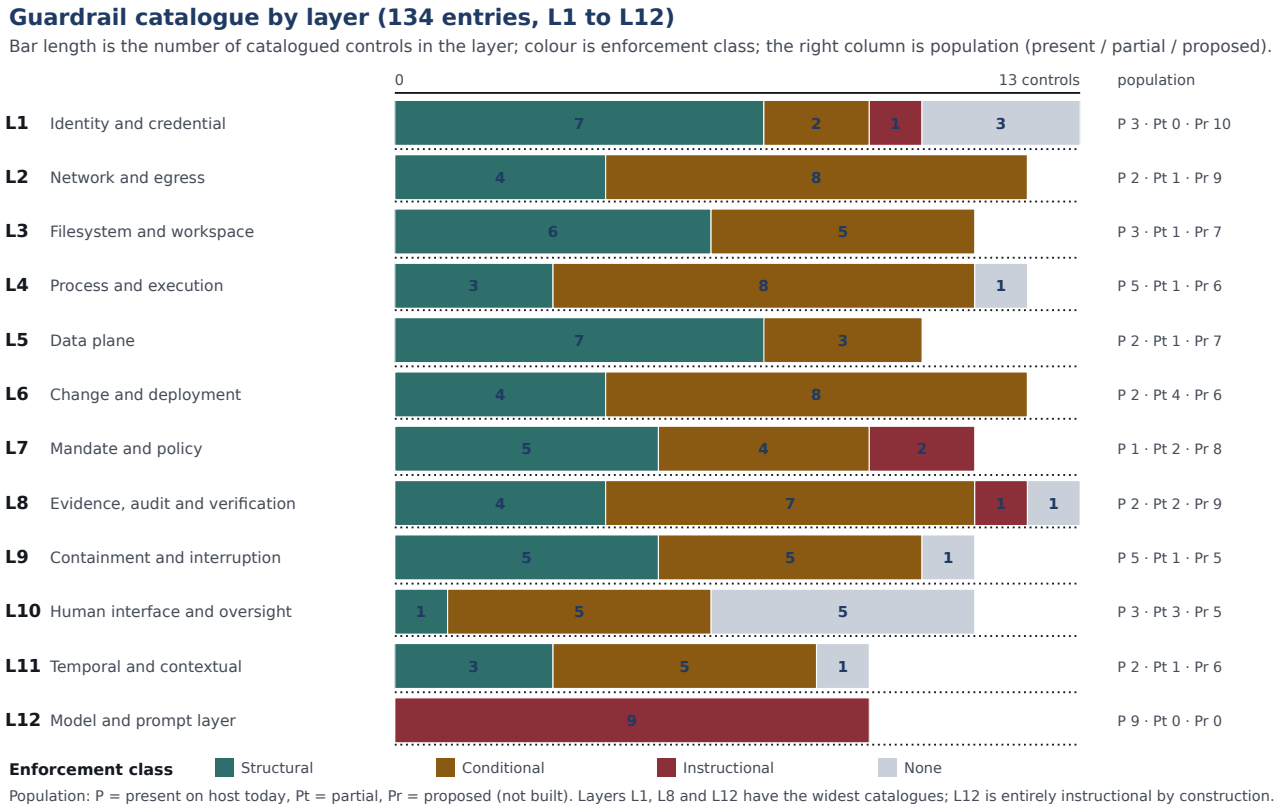

Figura D3-1. As doze camadas como uma pilha. O comprimento da barra é o número de controles catalogados na camada; a cor é a classe de enforcement; a coluna à direita é a população. O padrão se repete em quase todas as camadas, uma população presente dominada por entradas conditional e instructional, com a cobertura structural concentrada na coluna de propostas, e a L12 está presente, é numerosa e é inteiramente instructional por construção.

A L12 é deliberadamente catalogada por completo mesmo que todas as suas nove entradas sejam instructional, porque a análise que este catálogo existe para sustentar precisa mostrar, não apenas afirmar, quanto da postura atual do caso fundador repousa sobre uma camada que não consegue impedir nada de modo structural. Comparar a barra da L12 com as da L1 até a L11 na Figura D3-1 é a forma mais rápida de ver o argumento do D2 reafirmado como uma imagem.

1.2 A classificação de enforcement

Cada entrada carrega exatamente um `enforcement_class`, definido de modo idêntico onde quer que seja usado nesta base de conhecimento:

- **Structural.** A ação proibida não é alcançável pelo agente. Derrotá-la exige mudar a arquitetura, não mudar uma decisão. A conformidade do agente não é um termo no resultado.
- **Conditional.** Existe uma verificação em um caminho de código, avaliada em tempo de execução, mas o agente poderia alcançar o mesmo efeito por outro caminho, ou poderia alterar a condição contra a qual ela é verificada.
- **Instructional.** Diz-se ao agente o que fazer ou não fazer, em prosa que o modelo lê como contexto. Nada no runtime bloqueia a ação.
- **Nenhuma.** Nenhum controle de qualquer tipo.

A classificação é uma propriedade do mecanismo, não da intenção por trás dele nem do campo population. Uma entrada pode estar presente no host, bem construída, e ainda assim ser classificada `none` ou `instructional` se nada no runtime de fato impede a ação que ela nomeia. Nove das trinta e nove entradas presentes neste catálogo são classificadas `none` exatamente por esse motivo: `GR-L01-01`

conta de sistema operacional compartilhada, `GR-L01-09` root sem senha, `GR-L01-10` problema de credencial de metadados alcançável, `GR-L04-10` supressão de verificação de permissões habilitada por padrão, `GR-L08-13` o substrato de hook de transcrição não utilizado, `GR-L09-10` supervisão de reinício sempre ativo, `GR-L10-07` visualização de terminal ao vivo, `GR-L10-08` resposta automatizada a prompts interativos, e `GR-L11-05` disparo agendado sem supervisão. Cada uma é uma característica real e observada do host **`EV-003`** **`EV-004`** **`EV-028`**. Nenhuma das nove restringe qualquer coisa; várias removem ativamente uma restrição que de outro modo existiria. Catalogá-las como entradas de guardrail em vez de omiti-las é deliberado: um catálogo que lista apenas o que ajuda não pode ser usado para calcular uma lacuna de cobertura, e quatro das nove são elas próprias o mecanismo pelo qual um controle posterior e melhor projetado seria contornado, o que a Seção 6 usa diretamente.

1.3 O princípio do sinal emitido

Um guardrail que não emite sinal observável não pode ser avaliado e, portanto, não pode ser uma variável no desenho experimental. Isso é enunciado como uma restrição de projeto sobre o próprio catálogo, não como uma observação sobre o caso fundador. Cada entrada, presente ou proposta, carrega um campo `emitted_signal` que descreve o evento observável que o controle produz quando atua, quando é contornado, ou quando deixa de funcionar silenciosamente. Um guardrail proposto que não pode ser instrumentado é redesenhado para que emita um sinal, ou é excluído do catálogo. A consequência prática é que `emitted_signal` não é decorativo: é o que a métrica `MG-061` cobertura de emissão de sinal de controle conta, e uma entrada do catálogo com um campo de sinal vazio ou vazio de conteúdo é um defeito no registro, não uma descrição legítima de um controle silencioso.

> **POR QUE ESTE PRINCÍPIO É ENUNCIADO AQUI, E NÃO DEIXADO IMPLÍCITO**
>
> A arquitetura de referência proposta pelo D6 e cada métrica da subfamília `control effectiveness` do registro de métricas dependem de que isso valha para toda entrada adicionada ao catálogo em qualquer execução futura. Um guardrail é uma hipótese sobre o que interrompe um agente. Sem um sinal emitido, a hipótese não é testável, e um guardrail não testável é indistinguível, de fora, de um que nunca foi construído.

1.4 Como ler uma entrada

Cada entrada apresentada na Seção 2 carrega o conjunto completo de colunas definido em `registers/SCHEMAS.md`: o que o controle pretende e como funciona mecanicamente; a classe de enforcement; as classes de ação que ele vincula, por identificador `AC-nnn`; o nível de autonomia que torna defensável para essas classes de ação; a precondição que já precisa valer para que ele seja instalável; uma análise de bypass obrigatória, que declara especificamente como um agente, ou um atacante de posse da identidade do agente, o derrota ou o contorna; um modo de falha, e se o controle falha no modo fail-open (falha aberta) ou fail-closed (falha fechada); o sinal emitido e as métricas sensíveis a ele; seu custo operacional e a fricção que impõe ao trabalho legítimo; o serviço AWS e Azure que o implementaria; se está presente neste host hoje; e uma nota sobre por que é interessante como variável experimental.

Dois campos merecem uma nota de leitura. `fail_state` registra o que acontece quando o próprio controle quebra, não quando o agente o derrota; um controle fail-open degrada silenciosamente e um controle fail-closed degrada de forma ruidosa, e essa diferença é frequentemente mais consequente do que o comportamento do controle em operação normal. `bypass_analysis` é obrigatório e específico por projeto: uma análise de bypass que diga que um controle "poderia teoricamente ser contornado" não é aceitável neste registro, e cada entrada declara o caminho concreto, em geral a conta compartilhada, o root sem senha, ou a gravabilidade do arquivo do qual o próprio controle depende.

1.5 Estendendo o catálogo

O catálogo é aditivo, conforme a Seção 4 do `README.md`. Uma execução posterior que adicione uma entrada aloca o próximo número `GR-Lnn-nn` livre para a camada à qual ela pertence, preenche cada coluna de `registers/SCHEMAS.md`, incluindo uma análise de bypass real e um sinal emitido real, e nunca renumera ou exclui uma entrada existente; uma entrada que deixa de se aplicar é marcada `retired` em sua coluna `status`, nunca removida, porque documentos já renderizados a citam. Uma nova entrada que participe de qualquer condition set precisa ser adicionada explicitamente a `condition-sets.json`, porque a participação em um condition set não é inferida a partir da camada ou da population. Espera-se que este catálogo cresça substancialmente além de 134 entradas em execuções futuras; a estrutura de doze camadas foi projetada para absorver esse crescimento sem renumeração, porque a numeração de cada camada é independente da de todas as outras.

---

2. O catálogo por camada

2.1 L1. Identidade e credencial

Esta camada é o alicerce sobre o qual as outras onze são condicionadas, e a própria estrutura de dependências do catálogo diz isso diretamente: os campos `precondition` de nove outras camadas nomeiam a separação de identidade como algo que já precisa valer antes que o controle em questão faça sentido. A postura do caso fundador aqui é o exemplo mais claro, em todo o catálogo, de uma camada que está presente, é observada e, ao mesmo tempo, estruturalmente ausente. Uma única conta de sistema operacional executa toda sessão de agente, todo serviço supervisor e o backend do console, e essa conta possui root sem senha e sem restrições, além de alcançar o serviço de metadados da instância de nuvem sem restrição **`EV-003`** **`EV-033`**. Ambas são catalogadas como entradas presentes com classe de enforcement `none`, porque não há nada de parcial na ausência de enforcement delas; são a condição permanente que a análise de bypass de todas as demais camadas acaba citando.

Dez das treze entradas de L1 são propostas, e oito dessas dez são structural, o que torna esta a camada com a maior concentração de propostas structural do catálogo. Essa concentração não é incidental. Identidade de carga de trabalho por agente (`GR-L01-02`), privilégio mínimo por classe de ação (`GR-L01-03`), e credenciais efêmeras sem material de chave de longa duração em disco (`GR-L01-04`, `GR-L01-05`) são as quatro entradas que o D2 nomeia como a precondição para que quase tudo o mais no catálogo se torne structural em vez de conditional, porque um controle construído sobre uma identidade compartilhada e detentora de root herda a capacidade dessa identidade de contorná-lo.

Camada 1. Identidade e credencial.

GR-L01-01 Shared operating system account for all agents

INTENT	Provide one operating system identity under which every agent session, every supervising service and the console backend runs, so that shared context, credentials and tooling are reachable with no per-agent provisioning.
MECHANISM	All live agent sessions, the nine process-manager services and the console backend execute as a single uid. Session file modes are set explicitly to 0750 directories, 0640 records and 0600 tokens, but those modes exclude only group and other; the owner is the same uid for every agent.
ENFORCEMENT	none
BINDS	AC-016 AC-022 AC-023 AC-027 AC-053 AC-054
ENABLES	III
PRECONDITION	None; it is the default posture of a hand-provisioned single-tenant host. Undoing it requires per-agent accounts or containers, which the fleet supervision scripts and the shared context store would both have to be redesigned around.
BYPASS	There is nothing to bypass; this entry catalogues the absence of identity separation. Its consequence is that cross-session read, write, impersonation and deletion require no privilege escalation and no defect, only an ordinary tool call: the per-session bus token is a 0600 file readable by every agent, so one agent can authenticate as another on the inter-agent bus and type into a peer's live terminal. What the arrangement buys is operational: one credential set, one context store and one set of tools, which is why it persists.
FAILURE MODE	Not a failure mode but a standing condition. A defect in one agent, or a prompt injection reaching one agent, executes with the authority of every other agent and of the console that supervises them, and no artefact records which agent acted.
FAIL STATE	fail-open
EMITTED SIGNAL	none today; process accounting, file ownership and audit records all name the same principal, so no emitted signal distinguishes one agent from another. To be measurable it would have to carry a per-agent principal on every observable action, for example a distinct uid in kernel audit records or a distinct workload identity in cloud control plane logs.
METRICS	MG-040 MG-042 MG-063 MG-067
COST	low
FRICTION	low
AWS BINDING	No native equivalent; this is host identity arrangement. The corresponding AWS control is one IAM role per agent workload, assumed through IAM Roles Anywhere or a task role, with STS session tags naming the agent.
AZURE BINDING	No native equivalent; the corresponding Azure control is Microsoft Entra ID workload identity with one user-assigned managed identity per agent workload.
ON HOST	yes
EVIDENCE	EV-003 EV-023 EV-032 EV-004
RESEARCH VALUE	It is the baseline arm for every isolation experiment in this catalogue: holding the workload constant and varying only identity separation makes lateral reach and cross-session tamper directly measurable rather than argued.

GR-L01-02 Per-agent workload identity

INTENT	Give each agent session a distinct machine identity so that every action carries an attributable principal and authority can differ between agents.
MECHANISM	Launch each agent under its own account or container with a distinct uid, issue it a distinct workload credential exchanged for short-lived cloud role credentials through a federated trust, and stamp the agent name into the session name so the identity appears in every control plane log record.
ENFORCEMENT	structural
BINDS	AC-016 AC-022 AC-023 AC-027 AC-037 AC-042 AC-053
ENABLES	II
PRECONDITION	Agent processes must not share a uid and must not hold passwordless root; per-session filesystem isolation must exist so that credential material for one agent is not readable from another.
BYPASS	On this host it would be defeated at once, because the account holds passwordless root and can read another agent's credential file or assume its account outright. It becomes structural only when root is removed and the credential is held outside the agent's reach, in a broker running as a different uid or in attested hardware. A second, subtler route around it is conceptual: distinct identity without distinct authority buys attribution only, so an agent that can still do everything simply does it under its own name.
FAILURE MODE	A misconfigured trust mapping issues agent A the credential intended for agent B, producing silent privilege confusion that looks like correct attribution; a broker outage stalls every agent at launch.
FAIL STATE	fail-closed
EMITTED SIGNAL	A credential issue event per agent per session in the cloud audit trail, and the agent principal on every subsequent API call; observed off-host in the provider's control plane log rather than in anything the agent writes.
METRICS	MG-031 MG-040 MG-042 MG-064
COST	moderate
FRICTION	moderate
AWS BINDING	IAM Roles Anywhere or task roles with one role per agent, STS session tags carrying the agent name, CloudTrail for attribution.
AZURE BINDING	Microsoft Entra ID workload identity federation with one user-assigned managed identity per agent, Azure Monitor for attribution.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It is the variable that converts the question of which agent acted from unanswerable to answerable, and it permits a clean experiment on whether attribution alone, with authority held constant, changes claim divergence.

GR-L01-03 Least privilege permission set per action class

INTENT	Grant each agent only the permissions its mandated action classes require, so that an action outside the mandate fails at the credential rather than at the agent's judgement.
MECHANISM	Derive one policy per agent role from the action class register, a statement per permitted class with resource and region conditions, attach it as the only policy on that role, and cap the maximum with a permission boundary so the identity cannot widen itself.
ENFORCEMENT	structural
BINDS	AC-027 AC-038 AC-039 AC-040 AC-044 AC-048 AC-052
ENABLES	II
PRECONDITION	Per-agent identity must exist and a declarative action class policy must be expressed as data; without a distinct principal there is nothing to scope, and without a policy expressed as data the scoping has no source of truth.
BYPASS	Defeated today by the five standing administrator-class key pairs on disk and by the instance role the metadata service vends on demand: an agent that reaches any of those never touches its scoped role. Least privilege binds only when the scoped credential is the sole reachable one. There is also an internal defeat worth naming: one observed policy carries full identity-service rights, so a nominally scoped identity that retains the ability to create principals or attach policies can restore administrator in seconds, and only the permission boundary closes that.
FAILURE MODE	An over-tight policy denies a legitimate call mid-task and the agent, having no rollback, leaves partial state behind; an over-loose policy silently restores the previous posture while the register still records the intended scope.
FAIL STATE	fail-closed
EMITTED SIGNAL	An explicit access-denied record per out-of-mandate call, carrying the agent principal, the action and the resource, in the cloud control plane audit trail; this is the direct observable for boundary crossing attempts.
METRICS	MG-030 MG-032 MG-033 MG-042
COST	high
FRICTION	moderate
AWS BINDING	IAM identity policies with permission boundaries, service control policies at the organisation, AWS Verified Permissions with Cedar for application-level classes, CloudTrail for the denial record.
AZURE BINDING	Azure RBAC custom role definitions scoped per resource group, Azure Policy deny assignments as the boundary, Azure Monitor for the denial record.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It makes the gap between the prose mandate and the effective permission countable, turning access-exceeds-mandate from a judgement into a query, and it supports an experiment on whether tasks fail safely or messily when a credential refuses mid-task.

GR-L01-04 Ephemeral credentials with short session duration

INTENT	Bound the useful lifetime of any credential an agent holds so that a copied or exfiltrated credential expires before it can be reused.
MECHANISM	Replace static key pairs with role assumption returning session credentials valid for fifteen to sixty minutes, refreshed through a broker the agent calls rather than a file the agent reads, with the broker refusing refresh once the session's mandate window has closed.
ENFORCEMENT	structural
BINDS	AC-027 AC-037 AC-038 AC-039 AC-042 AC-043
ENABLES	II
PRECONDITION	No long-lived key material may remain on disk, and the refresh broker must run under an identity the agent cannot impersonate; otherwise the short-lived credential is simply the slower of two available paths.
BYPASS	Void on this host: five non-expiring key pairs sit in one file the agent owns, and the metadata service vends a fresh role credential on request, so expiry is defeated by re-minting rather than by attack. Even fully built, the control has a narrower reach than it appears: an agent that runs continuously refreshes continuously, so short duration bounds an exfiltrated credential and not a resident one. What it buys is revocability, because a trust policy change takes effect at the next refresh instead of requiring a key hunt.
FAILURE MODE	A refresh outage part way through a task fails an action that has already had partial effect; clock skew rejects valid credentials and presents as an authorisation problem rather than a time problem.
FAIL STATE	fail-closed
EMITTED SIGNAL	One credential issue event per refresh in the cloud audit trail, giving a directly countable lifetime distribution, plus an explicit expired-credential error whenever a stale credential is presented.
METRICS	MG-040 MG-042 MG-059
COST	moderate
FRICTION	low
AWS BINDING	AWS STS AssumeRole with an explicit session duration, IAM Roles Anywhere, Secrets Manager rotation for anything that must persist.
AZURE BINDING	Microsoft Entra ID managed identity tokens with short lifetimes, Azure Key Vault with short-lived shared access signatures and rotation policies.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It separates two risks usually conflated, credential theft and in-session blast radius, and predicts that varying duration alone moves neither realised blast radius nor boundary crossing success, which is a useful falsifiable claim.

GR-L01-05 No long-lived key material on disk

INTENT	Remove standing secrets from the filesystem the agent can read, so that possession of the agent's filesystem stops being possession of the estate.
MECHANISM	Delete the static cloud key pairs, the private key collection, the forge credential file and the secrets directory; replace each with a call to a broker that authenticates the caller's workload identity and returns a short-lived value in memory only, holding anything irreducibly static in a managed secret store behind that broker.
ENFORCEMENT	structural
BINDS	AC-004 AC-027 AC-036 AC-042 AC-043 AC-047 AC-039
ENABLES	II
PRECONDITION	A broker process running under an identity distinct from the agent, and removal of passwordless root, since a root-capable process can read another process's memory and environment.
BYPASS	Void by construction today: thirteen secret files, seven unencrypted private keys covering the estate's hosts, a forge credential file and five cloud key pairs are all owned by the agent's uid at mode 0600, which excludes every user except the one that matters. Even after removal the residual paths are real: an agent holding root reads the broker's process environment or intercepts the value in flight, so the control only closes when the broker is a different uid, root is gone and process tracing is restricted. What it buys immediately, even part built, is that a filesystem copy taken off the host stops being a complete estate compromise.
FAILURE MODE	A cached fallback file introduced to survive a broker outage reinstates exactly the exposure that was removed, and does so in a path nobody inventories; a broker outage otherwise stalls every credentialed action at once.
FAIL STATE	fail-closed
EMITTED SIGNAL	One secret retrieval event per use, naming the caller and the secret, in the secret store's audit log; on the host, the absence of readable key material is itself checkable by a scanner run under a different identity.
METRICS	MG-040 MG-042 MG-059
COST	moderate
FRICTION	moderate
AWS BINDING	AWS Secrets Manager and Systems Manager Parameter Store with IAM-scoped retrieval, instance credentials via STS only, Nitro Enclaves where key custody must be attested.
AZURE BINDING	Azure Key Vault with RBAC-scoped retrieval and managed identity authentication, Key Vault Managed HSM for custody.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It makes the reach of one compromised agent a computable quantity rather than a thought experiment, because lateral reach can be derived directly from the credential inventory before and after.

GR-L01-06 Just-in-time privilege elevation with expiry

INTENT	Replace standing administrative authority with a capability that is requested for a named action, decided by a different principal, and lapses automatically.
MECHANISM	Remove the blanket passwordless sudo entry; route privileged operations through a brokered request recording action class, target and justification, require an approver decision, then issue a time-boxed capability such as a brokered session or a narrowly scoped elevated role that expires without action.
ENFORCEMENT	structural
BINDS	AC-010 AC-029 AC-030 AC-031 AC-032 AC-033 AC-034
ENABLES	III
PRECONDITION	The agent account must not already hold passwordless root, and the fleet lifecycle scripts that create, reload, kill and delete agent units through privileged service management must be redesigned, because they depend on it.
BYPASS	Today there is nothing to elevate to: the account already holds unrestricted passwordless root, so the broker is skipped entirely. The removal is not cosmetic, which is the finding worth recording; four fleet lifecycle scripts depend on privileged service management, so an agent denied sudo loses the ability to manage its own fleet. After removal the residual route around it is that the agent still owns the wrapper scripts and unit files that the privileged helpers execute, so it can change what runs as root without holding root; that is why this control is only sound when paired with read-only mounts on the supervision code.
FAILURE MODE	An approver who is unreachable blocks privileged repair during an incident, which is the moment the pressure to restore standing access is highest; a broker that grants a wider capability than the request named reproduces standing privilege under an audit trail that reads as controlled.
FAIL STATE	fail-closed
EMITTED SIGNAL	An elevation request record and an approve or deny decision carrying approver identity, scope and expiry, plus a session recording for the elevated window; observed in the broker's log and in the brokered session history.
METRICS	MG-031 MG-042 MG-054 MG-071
COST	high
FRICTION	high
AWS BINDING	AWS Systems Manager Session Manager with session policies and session logging, temporary elevated access through STS with a permission boundary, IAM Identity Center permission sets with a session duration.
AZURE BINDING	Microsoft Entra Privileged Identity Management eligible assignments with activation approval and expiry, Azure Bastion for brokered host access.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It is the cleanest test of the cost side of the thesis, because it is the one control that reliably raises approval wait share and approval turnaround, so an experiment can price the autonomy tax of prior approval against the reduction in realised blast radius.

GR-L01-07 Separate identity for the verifier

INTENT	Ensure that whatever checks an agent's work does so under a principal the acting agent can neither assume nor silence.
MECHANISM	Run verification as a distinct workload identity holding read-only rights over the acting agent's outputs and write rights only to an evidence store the acting agent cannot write; deny the acting agent's role any permission over the verifier's compute, credentials or output.
ENFORCEMENT	structural
BINDS	AC-022 AC-023 AC-053 AC-054 AC-055
ENABLES	II
PRECONDITION	Per-agent identity, a separated evidence store, and removal of the shared account and of passwordless root; without those the verifier is the verified under another name.
BYPASS	On this host a verifier would share the uid of the agent it verifies, so it could read the verifier's own context, rewrite its instruction file, overwrite its output and amend the record afterwards; the adversarial review practised today is therefore self-review with extra steps. Separation becomes real only when the verifier's credential is unreachable from the acting agent's process. Even then one route remains open and cannot be closed by identity alone: the acting agent controls the artefacts under test, so it can shape what the verifier sees, which is why this must be paired with cross-checking against raw sources rather than against the agent's summary.
FAILURE MODE	If the verifier cannot run and the pipeline treats absence of a verdict as a pass, completion claims are accepted by default while the register still shows verification in place.
FAIL STATE	fail-closed
EMITTED SIGNAL	A verification record signed by the verifier principal carrying the claim, the evidence consulted and the verdict, written to a store the acting agent cannot modify; observed as a write event in that store's audit trail.
METRICS	MG-022 MG-023 MG-024 MG-067
COST	moderate
FRICTION	moderate
AWS BINDING	A separate IAM role for the verifier with write-once access to an S3 evidence bucket under Object Lock, and a bucket policy denying the actor role.
AZURE BINDING	A separate Entra managed identity with write access to immutable blob storage under legal hold, denied to the actor identity by RBAC.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The founding case measured 61.1 per cent factual divergence between agent self-report and independent evidence with no independent verifier present, so this is the natural treatment arm and it makes verification independence a computable ratio rather than an assertion.

GR-L01-08 Separate identity for the approver

INTENT	Make the principal that records an approval structurally distinct from the principal that performs the approved change.
MECHANISM	Hold approval state in a store writable only by an approver identity bound to a human authentication event with a hardware-backed factor, deny the actor's role any write to that record, and have the change path read the record rather than accept an asserted flag travelling with the request.
ENFORCEMENT	structural
BINDS	AC-006 AC-039 AC-044 AC-045 AC-056 AC-057
ENABLES	III
PRECONDITION	An approval record that is not a file on the agent's own filesystem, per-agent identity so the actor role can be denied, and an authentication factor for the approver that the agent cannot replay.
BYPASS	Today the approval gate's state and the agent's own mandate live in directories the agent owns, and the console treats authentication as authorisation with no role model at all, so a single session credential both performs and approves; an agent reaching that credential occupies both roles with no privilege change and no separation-of-duties event. Once the record is held under a different principal the technical route closes, but a social route remains: the agent authors the request text, so it can obtain a genuine approval for a misdescribed change. That is why this control is incomplete without approval requests that present evidence rather than the requester's summary.
FAILURE MODE	An approver identity outage blocks all gated change; a trust policy that quietly permits the actor to write the approval leaves the gate fully functional in appearance and entirely decorative in effect.
FAIL STATE	fail-closed
EMITTED SIGNAL	An approval record carrying approver principal, the hash of the subject artefact and a timestamp, plus an explicit denial event whenever an actor identity attempts to write it; the denial is the tamper signal.
METRICS	MG-031 MG-051 MG-055 MG-067
COST	moderate
FRICTION	high
AWS BINDING	A separate IAM principal writing approval objects to S3 with Object Lock or to a table denied to the actor role, IAM Identity Center for the human factor, CloudTrail for the record.
AZURE BINDING	A separate Entra identity writing to immutable blob storage, Entra Privileged Identity Management approval workflow for the human factor, Azure Monitor for the record.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It isolates separation of duties as a single variable and makes violations a simple count, which permits the sharp experiment of whether a gate whose record the actor can write performs measurably differently from no gate at all.

GR-L01-09 Passwordless root available to the agent account

INTENT	Allow every agent and every fleet helper script to manage service units, install packages and restart workloads with no interactive prompt, which is what makes unattended fleet supervision function at all.
MECHANISM	The sudo policy for the single agent account grants both unrestricted command execution as any user and unrestricted execution with no password, and the account is a member of the administrative group. The fleet lifecycle scripts that create a unit, reload an agent, kill an agent and delete an agent all depend on non-interactive privileged service management.
ENFORCEMENT	none
BINDS	AC-010 AC-021 AC-026 AC-029 AC-030 AC-031 AC-032 AC-054
ENABLES	III
PRECONDITION	n/a: this is the host's as-provisioned state and has no precondition. The dependency runs the other way, since removing it requires rewriting the four fleet lifecycle scripts that rely on non-interactive privileged service management.
BYPASS	There is nothing to bypass; the control is absent by construction. This entry is catalogued because it is the term that voids the others: every file mode, namespace check, allowlist and retention rule elsewhere in this catalogue is enforced by the kernel against a uid that can become root without a password. Any control in the filesystem, process, data, evidence or containment layers that rests on ownership, or on a code path inside a process the agent owns, is advisory while this holds. It is also load-bearing rather than incidental, so its removal is a redesign of fleet supervision rather than a configuration edit, which is precisely why it survives.
FAILURE MODE	n/a: there is no control to misconfigure. The operational failure it produces is that one defective or injected instruction reaches root on the host and, through the key material in the same home directory, the wider estate.
FAIL STATE	fail-open
EMITTED SIGNAL	none today; privileged invocations are written to the system authorisation log, but nothing correlates an invocation with the agent session or turn that issued it, and the agent can rewrite that log. To be measurable, each privileged invocation would have to carry a session identifier and land in a store the agent cannot write.
METRICS	MG-031 MG-042 MG-059 MG-060
COST	low
FRICTION	low
AWS BINDING	No AWS service manages a host sudo policy, so this is host configuration; the corresponding control is Systems Manager Session Manager with session policies replacing interactive elevation, on an instance profile that grants no host-level privilege.
AZURE BINDING	No Azure service manages a host sudo policy, so this is host configuration; the corresponding control is Azure Bastion access plus Entra Privileged Identity Management elevation in place of standing sudo.
ON HOST	yes
EVIDENCE	EV-003 EV-020 EV-032
RESEARCH VALUE	It is the founding case's dominant confound: any measured effect of a lower-layer control must be interpreted against it, so an experiment that removes it first and then reinstates it as a treatment isolates how much of the catalogue is enforcement and how much is convention.

GR-L01-10 Instance metadata credential issue reachable by agent processes

INTENT	Give workloads on the host a cloud identity with no provisioning and no rotation burden, which is the standard hosting pattern and is what lets co-located services call cloud interfaces unattended.
MECHANISM	The instance metadata service is reachable from an ordinary unprivileged process. Token-required mode is enforced, so an unauthenticated read is refused, but a token request succeeds on the first attempt from a plain shell and returns the metadata index, the instance identity document and a live role credential carrying an expiry.
ENFORCEMENT	none
BINDS	AC-027 AC-037 AC-038 AC-039 AC-040 AC-048
ENABLES	III
PRECONDITION	n/a: the path is open as provisioned and requires nothing to be true. Closing it requires either detaching the instance profile through the control plane or filtering the endpoint by owning uid, both of which need a substitute credential path for the co-located services that rely on it.
BYPASS	Nothing to bypass; the credential path is open by design. Token-required mode is a genuine control against a confused HTTP client performing request forgery, but an agent is a process that executes code, so it issues the token request itself and the enforcement is irrelevant to it. Two things the arrangement still buys are worth recording: the credential it vends carries an expiry, unlike the five static key pairs beside it on disk, and its use produces issue and call records on the provider side. What it costs is that even a perfect on-disk credential hygiene programme leaves a vending machine reachable from every agent, every subprocess and every script an agent writes.
FAILURE MODE	n/a as a control. The operational failure it produces is that a prompt injection reaching any agent process obtains a valid cloud identity in a single exchange, and the permission shape of that role was not determinable from the host, so the resulting blast radius is not knowable locally.
FAIL STATE	fail-open
EMITTED SIGNAL	none observable on the host; the credential fetch leaves no local record at all. Subsequent calls appear in the provider's control plane trail attributed to the instance role, but the agent's default identity is denied the read that would list audit trail configuration, so it cannot confirm it is being logged. Measurability would require a network-level record of metadata requests attributed to a process.
METRICS	MG-030 MG-040 MG-042 MG-059
COST	low
FRICTION	low
AWS BINDING	The metadata service itself; the corresponding control is detaching the instance profile, or a packet filter dropping traffic to the link-local address from every uid except a broker, with credentials reissued through IAM Roles Anywhere.
AZURE BINDING	The Azure Instance Metadata Service is the equivalent surface and issues managed identity tokens the same way; the corresponding control is removing the assigned managed identity and using Entra workload identity federation per process.
ON HOST	yes
EVIDENCE	EV-009 EV-033
RESEARCH VALUE	It is the clean counterexample to the claim that on-disk credential hygiene is sufficient: an experiment removing every key file while leaving the endpoint reachable should show lateral reach barely move, which is a strong result for the architectural-enforcement thesis.

GR-L01-11 Metadata service session hop limit and version enforcement

INTENT	Reduce the population of things that can obtain the instance's cloud identity to processes on the instance itself, and require any obtainer to hold a session token.
MECHANISM	Set the instance metadata options to token-required mode with a response hop limit of one, so a forwarded or containerised request cannot reach the endpoint. Token-required mode is already in force on this host, but the hop limit lives in the compute control plane and was not readable locally, so the pair is catalogued as not in place.
ENFORCEMENT	conditional
BINDS	AC-036 AC-037 AC-039
ENABLES	III
PRECONDITION	Control plane authority to modify the instance's metadata options, and no container runtime on the host that legitimately needs a hop count above one.
BYPASS	It does not bind the governed party at all. An agent executes on the instance, so its hop count is one by definition and the token request is a single command; the control binds a forwarding proxy or a compromised container, not an agent. It is further undone by the fact that the agent holds identity and compute rights in more than one account, so it can raise the hop limit through the control plane itself. Treating it as a control on agents is a category error, and cataloguing that error is the point of the entry.
FAILURE MODE	A hop limit of one breaks containerised workloads that legitimately need the credential, so it is commonly reverted, and the revert is invisible from the host because the setting is not locally readable.
FAIL STATE	fail-closed
EMITTED SIGNAL	none on the host; the setting is visible only as a control plane attribute and a change to it appears in the control plane audit trail. To be measurable at the agent layer it would need a per-request metadata access log, which the service does not provide.
METRICS	MG-033 MG-042 MG-059
COST	low
FRICTION	low
AWS BINDING	EC2 instance metadata options with tokens required and a put response hop limit of one, enforced fleet-wide by AWS Config rules or by a service control policy conditioned at launch.
AZURE BINDING	The Azure Instance Metadata Service exposes no hop limit; the nearest equivalent is removing the assigned managed identity and enforcing that with Azure Policy.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It is a valuable negative control in an experiment: a widely recommended hardening step that should show no movement at all in lateral reach for a code-executing agent, which distinguishes controls that bind this threat model from controls that bind a different one.

GR-L01-12 Credential inventory and rotation schedule

INTENT	Maintain an authoritative list of every credential the agent plane can reach, with an owner, a rotation interval and a last-rotated date, so that exposure has a known half life.
MECHANISM	A scheduled scanner enumerates key material on disk, the environment of every supervised process and the key metadata held by each cloud identity, writes the inventory to a store the agent cannot modify, and flags anything past its rotation interval; rotation itself is executed by a broker identity rather than by the agent.
ENFORCEMENT	instructional
BINDS	AC-004 AC-027 AC-037 AC-042 AC-047
ENABLES	III
PRECONDITION	A store the agent cannot write and a broker identity able to rotate without the agent's participation; otherwise the inventory is a map drawn by the party it describes.
BYPASS	The scanner runs as the same account it audits, so an agent can exclude a path, edit the inventory or rotate a key without recording it. It is also structurally blind in a way the present host demonstrates: model provider and third-party keys are injected into process environments rather than left in files, so a filesystem scan misses an entire credential class. Strictly it is a detective measure with no enforcement path, which is why it sits at the weakest class here. What it uniquely buys is that the count of standing credentials becomes a number, and that number is the input to every containment metric in the register.
FAILURE MODE	Silent under-enumeration: the inventory looks complete and is treated as complete while a whole class of credential, in a process environment or held only in memory, never appears in it.
FAIL STATE	fail-open
EMITTED SIGNAL	An inventory snapshot per run carrying a count and an age distribution per credential class, plus a diff event whenever a credential appears or disappears; observed in the inventory store. No such artefact exists on this host today.
METRICS	MG-042 MG-050 MG-064
COST	moderate
FRICTION	low
AWS BINDING	IAM credential reports and access key last-used data, AWS Config rules for key age, Secrets Manager rotation schedules, Security Hub for aggregation.
AZURE BINDING	Microsoft Entra ID application credential expiry reporting, Azure Key Vault rotation policies and near-expiry events, Microsoft Defender for Cloud recommendations.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It turns the privilege breadth index from a qualitative judgement into a computed value, which is the precondition for comparing two hosts, or one host before and after a change, on anything other than impression.

GR-L01-13 **Model provider credential scoped per agent**

INTENT	Give each agent its own model provider credential so that spend, rate limiting and revocation act on one agent rather than on the whole fleet.
MECHANISM	Issue one provider credential per agent session from a broker holding the parent credential, inject it into that session's environment at launch and revoke it at session end, so the agent never holds the parent and per-agent usage is separable at the provider.
ENFORCEMENT	conditional
BINDS	AC-014 AC-019 AC-036 AC-058
ENABLES	II
PRECONDITION	Per-agent identity and per-session environment isolation; where every supervised process environment is readable by one uid, per-agent scoping is nominal.
BYPASS	An agent reads any peer's credential straight out of another process's environment, and the observed host already carries provider keys for several providers across several supervised services in exactly that form. The fleet also pushes a shared provider credential to seven hosts deliberately, so revoking one agent revokes all of them, which is the opposite of scoping. Scoping becomes real only when a session's environment is unreadable from another session, which is the container or per-uid boundary again. What it still buys unenforced is per-agent attribution of model consumption, which is the only route to treating resource accounting as an independent signal rather than a self-report.
FAILURE MODE	A broker outage prevents session launch outright; a leaked parent credential silently restores fleet-wide reach while every per-agent record continues to look correctly scoped.
FAIL STATE	fail-closed
EMITTED SIGNAL	Per-agent usage and rate-limit events recorded at the provider, plus issue and revoke events per session at the broker; the provider-side record is produced off-host and is not written by the agent, which is what makes it independent.
METRICS	MG-006 MG-040 MG-048
COST	moderate
FRICTION	low
AWS BINDING	Where the model is consumed through Amazon Bedrock, one IAM role per agent with model invocation scoped by inference profile and per-role quotas; otherwise no native equivalent and the broker is implemented in the workload.
AZURE BINDING	Where the model is consumed through Azure OpenAI, one Entra managed identity per agent with RBAC on the deployment; otherwise no native equivalent and the broker is implemented in the workload.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Provider-side usage is one of very few signals on this host that the agent does not author, so per-agent scoping converts it into a genuine independent oversight channel and makes oversight signal independence non-trivially computable.

2.2 L2. Rede e saída

Esta camada governa o que os processos de um agente conseguem alcançar em uma rede, nas duas direções. A postura de entrada (inbound) do caso fundador é comparativamente sólida: um proxy reverso é o único ponto de entrada público e encerra o TLS com um piso de protocolo moderno, ambos presentes e conditional `EV-008`. A postura de saída (outbound) não é sólida de forma alguma. Nenhum firewall de host está ativo e a saída (egress) é irrestrita; nove das doze entradas, incluindo negação por padrão da saída, uma lista de permissão de destinos e o bloqueio do acesso do processo do agente ao endpoint de metadados de nuvem, são propostas `EV-009`. A alcançabilidade do endpoint de metadados, registrada como estruturalmente ausente na L1, e a saída irrestrita registrada aqui se compõem diretamente: um processo de agente que consegue alcançar o serviço de metadados e também consegue alcançar qualquer destino na internet pode cunhar a identidade de nuvem da instância e

exfiltrar o que quer que obtenha, na mesma ação, sem nada nesta camada posicionado para perceber qualquer uma das duas metades acontecendo.

Quatro das entradas propostas da camada são structural, e todas as quatro dependem da mesma precondição, a de que a saída seja negada por padrão para que uma lista de permissão tenha algo do qual ser exceção. A segmentação entre níveis de agente e a ausência de acesso lateral de rede entre espaços de trabalho de agentes (`GR-L02-08` , `GR-L02-09`) são o análogo, no nível de rede, da identidade por agente: nenhuma das duas faz sentido enquanto todo agente já compartilha a conta que a L1 deixa não separada.

Camada 2. Rede e saída.

`GR-L02-01` Reverse proxy as the sole public entry point

INTENT	Concentrate all public access to the agent plane at one terminating hop, so that transport security, authentication routing and header handling exist in exactly one place.
MECHANISM	nginx terminates the public listener for a single server name, redirects plaintext to the encrypted listener, and proxies onward to two backends bound on the loopback interface; a catch-all location forwards everything to the console backend and a prefixed location forwards to the operator authentication service. The proxy injects a marker header so the backend can recognise requests that arrived through it.
ENFORCEMENT	conditional
BINDS	AC-033 AC-034 AC-035 AC-048
ENABLES	II
PRECONDITION	A packet filter or cloud security group must actually prevent direct connection to the backend listeners; without one, sole entry point describes an intention rather than a topology.
BYPASS	It is the sole entry point only for the two services that bind loopback. Observed on this host, four other listeners bind on all interfaces or on the host's primary address, including the agent runtime that exposes remote shell execution and agent spawning as model-callable tools, and with no packet filter and unrestricted routing any peer on the local network reaches them without touching the proxy. A second, documented gap: the backend's machine-token routes were believed to be unproxied, but the single catch-all location forwards them, so the containment for the most dangerous internal routes is the token alone and not the topology. From the agent's own side the proxy constrains nothing at all, because it governs inbound traffic only.
FAILURE MODE	A configuration error or certificate lapse takes the entire operator interface offline, which also removes the intervention affordance; a location added without an exclusion silently publishes an internal route.
FAIL STATE	fail-closed
EMITTED SIGNAL	An access record per request with method, path, status and forwarded client address, plus an error log; both are files on the host under the same account as the agent, so the signal is observable but not tamper-resistant.
METRICS	MG-031 MG-047 MG-063
COST	low
FRICTION	low
AWS BINDING	An Application Load Balancer or CloudFront distribution as the only ingress, with security groups admitting the backend only from that load balancer, and AWS WAF applied at the same edge.
AZURE BINDING	Azure Application Gateway or Azure Front Door as the only ingress, with network security groups restricting backend reachability and Azure Web Application Firewall at the edge.
ON HOST	yes
EVIDENCE	EV-006 EV-008 EV-016
RESEARCH VALUE	A concrete instance of a control that is architecturally correct at the component and empirically incomplete at the topology, which makes it a compact worked example of why structural is a property of the whole path and not of the proxy.

GR-L02-02 TLS termination with modern protocol floor

INTENT	Ensure that operator traffic to the agent plane, which includes terminal streams carrying credentials and keystrokes into live agent sessions, cannot be read or altered in transit.
MECHANISM	The public server block pins the two current TLS versions, serves a wildcard certificate renewed by an ACME client, sets a one-year strict transport security header including subdomains, and adds content-type sniffing prevention, frame denial and a referrer policy; plaintext requests are redirected. The global block still lists two obsolete protocol versions, which the server block overrides for this service only.
ENFORCEMENT	conditional
BINDS	AC-033 AC-034 AC-041 AC-048
ENABLES	II
PRECONDITION	Certificate material and the renewal path must sit outside the agent's write reach, otherwise the protocol floor is a setting the governed party maintains for itself.
BYPASS	It protects the operator's transport, not the agent's behaviour, and the agent can rewrite the server block and reload the proxy with the root it already holds; a downgrade would be visible only in a browser indicator. The inherited global protocol list is the more interesting weakness: any new server block that omits its own directive silently inherits an obsolete floor, so the effective posture depends on a per-block repetition rather than on a single enforced default. What the control genuinely buys is that the terminal WebSocket, which is the channel that types into running agents, is not exposed in clear text on the wire.
FAILURE MODE	Certificate expiry breaks the operator console entirely, removing the intervention affordance at the moment it may be most needed, and the strict transport policy then deliberately prevents a plaintext fallback.
FAIL STATE	fail-closed
EMITTED SIGNAL	Handshake failures and renewal outcomes in the proxy and ACME client logs; critically, no signal at all is emitted when the protocol floor itself is edited, which is the observability gap rather than the cryptography.
METRICS	MG-047 MG-050 MG-063
COST	low
FRICTION	low
AWS BINDING	AWS Certificate Manager with an explicit TLS security policy on the load balancer, or a CloudFront minimum protocol version, with configuration drift caught by AWS Config.
AZURE BINDING	Azure Front Door or Application Gateway TLS policy with a minimum protocol version, certificates sourced from Azure Key Vault, drift caught by Azure Policy.
ON HOST	yes
EVIDENCE	EV-008
RESEARCH VALUE	A clean, small worked example of nominal posture diverging from effective posture, since the global setting and the serving setting disagree, which is exactly the quantity the posture divergence metric is meant to count.

GR-L02-03 Loopback-only bind for sensitive local services

INTENT	Keep services holding privileged capability unreachable from the network by binding them to the loopback interface, so that reaching them requires already being on the host.
MECHANISM	The console backend, the local knowledge search server, the resource accounting service and the operator authentication service each bind the loopback address; the reverse proxy is the only route to the two that are meant to be publicly reachable.
ENFORCEMENT	conditional
BINDS	AC-016 AC-035 AC-042 AC-043
ENABLES	II
PRECONDITION	A packet filter as a second layer, because a bind scope is chosen by the process itself and can be changed by anyone who can edit the code or restart the service.
BYPASS	Partial by direct observation: alongside four loopback binds there are four listeners on all interfaces or on the host's primary address, and the most consequential of them is the agent runtime that offers remote execution and agent spawning as model-callable tools with a host allowlist but no command allowlist. There is no packet filter and no verified network boundary, so those four are reachable from the local network. Separately, the agent owns both the service code and the process manager, so changing a bind address and restarting is an ordinary action requiring no privilege it does not already hold. What loopback binding still buys is that the backend's machine-token routes cannot be reached directly from the network, only through the proxy.
FAILURE MODE	A service that falls back to binding all interfaces after a configuration edit exposes itself silently, because nothing on this host alerts on the appearance of a new externally reachable listener.
FAIL STATE	fail-open
EMITTED SIGNAL	none today; bind scope is observable only by inspecting listening sockets on demand. A measurable form would be a periodic listener inventory, diffed against an expected set and emitted to a store the agent cannot write.
METRICS	MG-031 MG-040 MG-050
COST	low
FRICTION	low
AWS BINDING	VPC security groups with no inbound rule other than from the load balancer, network ACLs at the subnet boundary, and PrivateLink for anything that must be shared between tiers.
AZURE BINDING	Network security groups with default-deny inbound at subnet and interface level, and Azure Private Link for shared services.
ON HOST	partial
EVIDENCE	EV-006 EV-008 EV-009
RESEARCH VALUE	It offers a natural within-host comparison: four services with the control and four without, under the same account and the same code ownership, so reachability differences are attributable to bind scope alone.

GR-L02-04 Host firewall with default deny inbound

INTENT	Refuse inbound connections to anything not deliberately published, so that a service's own bind choice is not the only thing between it and the network.
MECHANISM	Load an nftables rule set with a default-deny input policy, accept established and related flows, admit the public listeners and the administrative access path, and drop everything else; hold the rule set in configuration rather than in interactive commands so it is reproducible and reviewable.
ENFORCEMENT	conditional
BINDS	AC-012 AC-035 AC-042 AC-043
ENABLES	II
PRECONDITION	An accurate inventory of required listeners, because four services currently bind broadly and would break, and removal of the agent's root, without which the rule set is advisory.
BYPASS	The governed party holds passwordless root, so it flushes the table or inserts an accept rule with one command and nothing durable records that it did. It is also aimed at the less pressing direction here, since the dominant risk on an agent host is outbound rather than inbound. Against an external or lateral attacker it is genuinely structural, because the kernel drops the packet before any application sees it. The version that the governed party cannot reach is the same rule expressed as a cloud security group, which is why the provider binding matters more than the host binding for this entry.
FAILURE MODE	A rule set applied without an allow rule for the administrative path locks the operator out of the host entirely; a rule set that fails to load at boot leaves everything open while the configuration on disk still reads as enforced.
FAIL STATE	fail-closed
EMITTED SIGNAL	Per-chain drop counters and log rules per rejected flow, plus a rule set version and hash that can be attested at boot; no such signal exists today because no filter is loaded.
METRICS	MG-031 MG-040 MG-062
COST	low
FRICTION	moderate
AWS BINDING	VPC security groups, which are stateful and cannot be altered from inside the instance without control plane credentials, network ACLs at the subnet, and AWS Network Firewall for centrally managed rules.
AZURE BINDING	Network security groups at subnet and interface level with default-deny inbound, and Azure Firewall for centrally managed policy.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It is the catalogue's clearest illustration of the difference between an in-host and an out-of-host control against an identical threat, since the same rule expressed as a security group is not reachable by the governed party at all.

GR-L02-05 Deny by default egress

INTENT	Prevent an agent reaching any network destination that has not been deliberately permitted, so that exfiltration and unplanned external effect require a policy change rather than a decision.
MECHANISM	Place the agent host in a subnet with no default route to the internet, provide reachability only through a firewall endpoint or forward proxy that applies a destination policy, and route cloud service access through private endpoints so that a blanket deny does not break the control plane the workload depends on.
ENFORCEMENT	structural
BINDS	AC-004 AC-036 AC-042 AC-043 AC-047 AC-048 AC-050
ENABLES	II
PRECONDITION	Private endpoints for the cloud services the workload genuinely needs, and a settled list of model provider and package registry destinations, otherwise the agent plane stops working on its first turn.
BYPASS	It is the one control in this layer the governed party cannot defeat from inside the host, because no amount of root creates a route that does not exist. Two real routes around it remain. First, the agent holds administrator-class credentials in four cloud accounts, so it can edit the route table or the firewall policy through the control plane, which makes this control durable only in combination with least privilege at the identity layer. Second, any permitted destination that is itself general purpose, a package registry, a source forge, the model provider endpoint, is an exfiltration channel, so denial by default without content awareness relocates the problem rather than removing it.
FAILURE MODE	An unlisted destination fails a task part way with a network error the agent may misdiagnose and route around by other means; an outage at the single egress point stops the entire fleet at once.
FAIL STATE	fail-closed
EMITTED SIGNAL	One allow or deny record per connection attempt with destination, port and requesting workload, produced at the firewall or proxy. This is the richest oversight signal in the catalogue precisely because it is generated off-host and cannot be edited by the party it describes.
METRICS	MG-032 MG-033 MG-040 MG-041
COST	high
FRICTION	high
AWS BINDING	Private subnets with no internet gateway route, AWS Network Firewall or a forward proxy behind NAT with a domain allowlist, VPC endpoints for cloud services, Route 53 Resolver DNS Firewall alongside.
AZURE BINDING	Azure Firewall with application rules and forced tunnelling through user-defined routes, Private Link and Private Endpoint for platform services.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The single highest-value experimental variable in this layer, because toggling it simultaneously changes containment and creates an agent-unwritable event stream, so one manipulation moves both a containment metric and an oversight-independence metric.

GR-L02-06 Destination allowlist for agent egress

INTENT	Constrain the set of external destinations an agent may reach to those its mandate requires, and make additions a reviewed change rather than an in-session decision.
MECHANISM	A domain and address allowlist enforced at the forward proxy or firewall, expressed per agent role rather than per host, naming the model provider, the package registries, the source forge and the estate hosts each role legitimately needs; anything else is refused and recorded with the requesting workload.
ENFORCEMENT	conditional
BINDS	AC-004 AC-036 AC-047 AC-048 AC-050
ENABLES	II
PRECONDITION	Per-agent identity presented to the enforcement point, by client certificate or authenticated proxy credential; without it the list cannot be per role and degenerates to a per-host union.
BYPASS	On a single-account host the proxy sees one client, so the effective allowlist is the union of every agent's needs and the least constrained role sets the bound for all the others; that is a structural consequence of the identity layer, not a configuration oversight. Beyond that, any permitted general purpose destination is a tunnel: a source forge permitted for pushes is also a place to write arbitrary data, and a package registry permitted for installs is also a fetch channel for arbitrary code. Domain-level matching is further weakened wherever permitted names resolve onto shared infrastructure. What it still buys is that destination cardinality becomes a small, countable, reviewable number rather than the whole internet.
FAILURE MODE	An over-broad entry, typically a wildcard covering a large provider, restores effectively unrestricted egress while the policy document continues to read as restrictive.
FAIL STATE	fail-closed
EMITTED SIGNAL	A per-request allow or deny record naming the destination and the rule matched, plus a periodic count of distinct destinations reached per agent, which is directly the egress cardinality measure.
METRICS	MG-032 MG-033 MG-041
COST	moderate
FRICTION	high
AWS BINDING	AWS Network Firewall domain list rule groups, Route 53 Resolver DNS Firewall domain lists, or a proxy on a NAT instance authenticating each agent identity separately.
AZURE BINDING	Azure Firewall application rule collections with fully qualified domain name filtering and FQDN tags, with Azure Firewall Premium where per-identity inspection is required.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It makes egress cardinality a directly manipulable variable and exposes a dependency worth measuring, namely that per-agent network policy is unattainable without per-agent identity, which links the network layer to the identity layer empirically rather than rhetorically.

GR-L02-07 DNS egress control and resolution logging

INTENT	Record and constrain what an agent attempts to resolve, capturing destination intent even where the connection itself is blocked, tunnelled or never completed.
MECHANISM	Force all name resolution through a controlled resolver, apply domain rule groups blocking non-allowlisted and known-bad zones, export a query record per resolution with the requesting workload, and deny outbound resolution to any other resolver including encrypted transports.
ENFORCEMENT	conditional
BINDS	AC-036 AC-040 AC-042 AC-047
ENABLES	II
PRECONDITION	Egress control must already block direct resolution traffic and encrypted DNS to third parties, otherwise the controlled resolver is one option among many; the host currently runs a local stub resolver with an unrestricted default route.
BYPASS	An agent with unrestricted egress simply queries a different resolver, uses encrypted DNS to a general purpose provider, or omits resolution entirely by connecting to a literal address. It also holds full rights over the estate's public DNS service in at least one account, so where a name is inconvenient it can create the record it wants instead of evading the check. What resolution logging buys uniquely, and nothing else in this catalogue provides, is intent: it captures the destination an agent tried to reach even when the connection never succeeded.
FAILURE MODE	A resolver outage stops all outbound work at once; an over-broad block list breaks legitimate estate resolution and induces an agent to hardcode addresses, which then evades the logging permanently and silently.
FAIL STATE	fail-closed
EMITTED SIGNAL	One query record per resolution carrying the name, the response action and the requesting source, exported off-host; it is a leading indicator, available before the connection attempt rather than after it.
METRICS	MG-032 MG-041 MG-061
COST	low
FRICTION	low
AWS BINDING	Route 53 Resolver DNS Firewall with domain lists and query logging delivered to CloudWatch Logs or S3.
AZURE BINDING	Azure DNS Private Resolver with Azure Firewall acting as DNS proxy, and DNS query logging into Azure Monitor.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It supplies a leading measure of boundary-crossing intent that is independent of whether the crossing succeeded, which is precisely the distinction the metric register draws between attempt rate and success rate and which no present signal on this host can separate.

GR-L02-08 Segmentation between agent tiers

INTENT	Prevent an agent operating on a low-risk tier from reaching hosts or services belonging to a higher-risk tier.
MECHANISM	Place the orchestration plane, the development workloads and the production workloads in separate subnets, permit only the specific flows each tier requires through security groups that reference each other rather than address ranges, and require a brokered session for any cross-tier action instead of a direct connection.
ENFORCEMENT	structural
BINDS	AC-042 AC-043 AC-044 AC-045 AC-051 AC-052
ENABLES	II
PRECONDITION	A tier assignment per host that reflects real blast radius rather than history, a broker for the flows that must legitimately cross, and per-agent identity so that agents co-located on the orchestration host do not all inherit the union of tier reach.
BYPASS	The governed party cannot route around a security group from inside the instance, which is what makes this structural, but it can edit the group through the cloud control plane, and it holds administrator rights in four accounts. Segmentation is therefore only as strong as the least privilege applied to the identity that can rewrite it, which is the recurring dependency across this catalogue. On the present host segmentation is absent in effect rather than merely weak: private keys for every estate host sit in the agent's home directory and opening a shell to another host is a routine, exercised action visible in the live process list.
FAILURE MODE	A permissive rule added during an incident and never removed leaves the topology reading as segmented while the flow it opened persists indefinitely.
FAIL STATE	fail-closed
EMITTED SIGNAL	Flow records for accepted and rejected connections between tiers, plus configuration change events on the rules themselves; both are produced by the provider rather than by the host, so neither is written by the agent.
METRICS	MG-038 MG-039 MG-040
COST	moderate
FRICTION	moderate
AWS BINDING	Separate VPCs or subnets with security groups referencing each other, Transit Gateway route tables controlling reachability, VPC endpoints, and Session Manager for brokered cross-tier access.
AZURE BINDING	Virtual network segmentation with network security groups and Azure Firewall between spokes, and Azure Bastion for brokered cross-tier access.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It permits realised blast radius to be measured as a function of topology while agent behaviour is held constant, which is the only way to attribute an incident's reach to architecture rather than to the agent's choices.

GR-L02-09 No lateral network access between agent workspaces

INTENT	Ensure one agent cannot reach another agent's runtime surfaces over the network, so that compromise of one session does not become control of the fleet.
MECHANISM	Give each agent session its own network namespace or container with no route to its peers, expose only an outbound path through the controlled egress point, and force all inter-agent communication through a broker that authenticates both ends rather than trusting a shared file.
ENFORCEMENT	structural
BINDS	AC-014 AC-015 AC-016 AC-035 AC-058
ENABLES	II
PRECONDITION	Per-session process isolation and per-agent identity; a network control cannot separate two processes that already share a network namespace.
BYPASS	On this host the control is not merely bypassable, it is inapplicable: every agent session is a pane in one process tree on one host, sharing one network namespace and one uid, so lateral movement requires no network path at all. The existing inter-agent bus does authenticate with a per-session token, and that token is a file readable by every agent, so impersonating a peer on the bus is a file read rather than a network attack. Network separation only acquires meaning after separation at the process and identity layers, which makes this a dependent control and worth cataloguing chiefly for that ordering result.
FAILURE MODE	Any service every namespace still needs, the message broker or the terminal multiplexer, becomes the new lateral path and concentrates the risk the segmentation was meant to spread.
FAIL STATE	fail-closed
EMITTED SIGNAL	Connection attempts between agent namespaces, recorded and rejected at the namespace boundary or at the broker; today no such signal can exist because there is no boundary to attempt to cross.
METRICS	MG-038 MG-040 MG-062
COST	high
FRICTION	moderate
AWS BINDING	One task or instance per agent with its own security group and no group-to-group rules, and PrivateLink for the broker.
AZURE BINDING	One container group or virtual machine per agent with its own network security group, and Private Link for the broker.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Its precondition analysis is itself the finding: the experiment it supports demonstrates that network isolation between agents is unattainable without identity and process isolation first, which is a concrete defence-in-depth ordering result rather than a general appeal to layering.

GR-L02-10 Private endpoints for cloud service access

INTENT	Reach cloud services over private network paths so that a default-deny egress policy does not force a general internet exception for the control plane the workload needs.
MECHANISM	Provision interface and gateway endpoints for the services the workload actually uses, attach endpoint policies restricting the actions and resources reachable through them, and remove the public route so those services are reachable only through the endpoint.
ENFORCEMENT	structural
BINDS	AC-027 AC-036 AC-038 AC-039
ENABLES	II
PRECONDITION	Deny-by-default egress must be in force, otherwise the public service endpoint remains reachable and the private one is merely an additional path rather than the only one.
BYPASS	Without a blocked public route the agent calls the public endpoint and the endpoint policy is never evaluated. Even with the route removed, the agent holds credentials for five separate accounts, and an endpoint policy in one account does not constrain a call made with a different account's credential over whatever general egress remains permitted. The control's real contribution is enabling: it removes the strongest operational objection to denying egress by default, namely that the workload must reach the internet in order to reach its own cloud services.
FAILURE MODE	A missing endpoint for one service surfaces deep inside a task as an opaque timeout rather than a clear denial; an over-permissive endpoint policy restores full service access while the topology still reads as private.
FAIL STATE	fail-closed
EMITTED SIGNAL	Endpoint policy denials and per-endpoint flow records in the provider's logs; on the host itself nothing is emitted, which is the point, since the record is produced where the agent cannot reach it.
METRICS	MG-030 MG-031 MG-041
COST	moderate
FRICTION	low
AWS BINDING	VPC interface endpoints with endpoint policies, gateway endpoints for object storage, and condition keys restricting calls to a named endpoint.
AZURE BINDING	Private Endpoint and Private Link with private DNS zones, plus service endpoint policies where a service supports them.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It is the enabling dependency for the strongest control in the layer, so an experiment can quantify how much of the friction usually attributed to egress denial is in fact attributable to the absence of private paths.

GR-L02-11 Outbound rate limiting per agent identity

INTENT	Bound the speed at which an agent can produce external effect, so that a runaway or hostile loop is throttled to within human intervention latency.
MECHANISM	Apply per-identity connection and request rate limits at the egress enforcement point, with separate budgets per destination class and a burst allowance small enough that a repeated action trips the limit well before it completes at scale; exceeding the budget queues or refuses explicitly rather than dropping silently.
ENFORCEMENT	conditional
BINDS	AC-036 AC-046 AC-047 AC-050
ENABLES	II
PRECONDITION	Per-agent identity visible at the enforcement point; on a shared account the budget is per host, so the busiest agent starves every other and the limit becomes an availability problem rather than a control.
BYPASS	The governed party does not control the enforcement point, so the ceiling itself holds; the control is nevertheless aimed at the wrong quantity for the actions that matter most here. Changing one public DNS record, sending one message to a member of the public, or deleting one log group is a single request comfortably inside any sane budget. Rate limiting therefore contains fan-out incidents such as a notification storm or a bulk send, and does nothing whatever about single irreversible acts. In practice it is further diluted by an agent that spreads one effect across several permitted destination classes.
FAILURE MODE	A budget set below legitimate peak produces intermittent, non-deterministic failures, which is the hardest class for an agent to diagnose and the most likely to prompt a durable workaround that evades the limit thereafter.
FAIL STATE	fail-closed
EMITTED SIGNAL	Throttle events per identity and destination class, carrying the budget and the observed rate; useful as a leading indicator of loop behaviour before the loop produces external effect at scale.
METRICS	MG-002 MG-045 MG-046
COST	moderate
FRICTION	moderate
AWS BINDING	AWS WAF rate-based rules at the egress proxy, Network Firewall connection limits, and API Gateway usage plans where calls are brokered.
AZURE BINDING	Azure Firewall Premium rate limiting, and Azure API Management rate-limit-by-key policies where calls pass through a broker.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It is the only control in the catalogue that directly manipulates the latency margin, so it operationalises the supervision viability claim that intervention is meaningful only where harm completion time exceeds intervention latency.

GR-L02-12 **Block agent process access to the cloud metadata endpoint**

INTENT	Remove the agent's ability to mint the instance's cloud identity, leaving only credentials issued to it deliberately.
MECHANISM	Drop traffic to the link-local metadata address from every uid except a small broker account using an owner-matched packet filter rule, or detach the instance profile entirely and issue workload credentials through a federated exchange, with the broker handing the agent only a scoped short-lived credential.
ENFORCEMENT	conditional
BINDS	AC-027 AC-037 AC-038 AC-039 AC-040
ENABLES	II
PRECONDITION	An alternative credential path for the co-located services that legitimately need one, and either an owner-matched filter with the agent denied root, or control plane authority to detach the profile.
BYPASS	As a host filter rule it is defeated by a command the agent already holds the right to run, and the removal leaves no durable record anywhere. Detaching the instance profile is the variant the governed party cannot undo from inside the host, and it is strictly better because it removes the credential rather than the path to it. In both variants the residual gap is the same and is decisive: five static key pairs already sit on disk and never traverse the metadata endpoint, so this control changes nothing measurable unless it is paired with removing on-disk key material.
FAILURE MODE	Blocking the endpoint breaks any co-located service depending on the instance role, often as a long retry timeout rather than a clear error, which makes the cause hard to attribute and the revert tempting.
FAIL STATE	fail-closed
EMITTED SIGNAL	Dropped-packet counters for the metadata address broken down by owning uid, and, more durably, the disappearance of instance-role-attributed calls from the cloud control plane trail, which is an off-host observable.
METRICS	MG-033 MG-040 MG-042
COST	low
FRICTION	moderate
AWS BINDING	Detaching the instance profile, or an owner-matched nftables rule on the link-local address, with credentials reissued through IAM Roles Anywhere or an OIDC federation.
AZURE BINDING	Removing the assigned managed identity from the virtual machine and adopting Entra workload identity federation; the Azure metadata address can be filtered in the same way.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Paired with the removal of on-disk key material it separates two credential paths that are habitually treated as one problem, and predicts that closing either alone leaves lateral reach essentially unchanged.

2.3 L3. Sistema de arquivos e espaço de trabalho

Esta camada governa o que o agente pode ler, escrever ou ter isolado dele em disco. O caso fundador executa uma lista de permissão de realpath que faz a mediação do acesso do console ao sistema de arquivos, e uma raiz de arquivos com namespace por sessão, ambas presentes e ambas conditional; a lista de permissão canonicaliza antes de comparar e, como implementação, é sensível a separadores, mas as raízes que ela permite são os dois diretórios home que contêm efetivamente tudo no host, ambos graváveis, e um conjunto documentado de superfícies de escrita nunca a consulta `EV-018` `EV-023`. Essa é a lição central da camada: um controle conditional pode ser implementado corretamente e ainda assim falhar como controle, porque a correção da verificação e a adequação do escopo são propriedades independentes.

As duas entradas que esta camada compartilha com a discussão de mandato da L7 são as que o D2 trata como a forma mínima viável de um agente estruturalmente incapaz de ampliar sua própria

autoridade: a negação de acesso de escrita aos próprios arquivos de instrução do agente e às suas próprias configurações de permissão (GR-L03-07 , GR-L03-08). Ambas são propostas. Hoje, os arquivos de instrução de um agente e seu arquivo de configurações de permissão são graváveis pela mesma conta que os lê como autoridade, o que torna todo controle instructional na L12 revisável pela própria parte que ele deveria governar EV-032 .

Camada 3. Sistema de arquivos e espaço de trabalho.

GR-L03-01 Realpath allowlist on console-mediated filesystem access

INTENT	Confine every filesystem operation performed through the console's file interface to a declared set of roots, with read and write scope stated separately per root.
MECHANISM	A single resolver canonicalises the requested path with realpath and asserts that the result lies inside one of four configured roots, two writable and two read-only. Write, upload, rename and delete additionally require the resolved root to be writable, rename requires it at both ends, delete requires an explicit confirmation field, and path length and payload size are bounded by schema validation before the resolver runs.
ENFORCEMENT	conditional
BINDS	AC-022 AC-023 AC-024 AC-028
ENABLES	II
PRECONDITION	That the interface is the only route to the filesystem for the governed party, which is false here, and that the writable roots do not themselves contain the governance artefacts, which is also false.
BYPASS	The agent does not use the interface. Its own file tools and any shell command operate on the same filesystem under the same uid with no resolver anywhere in the path, so the allowlist bounds the console's clients rather than the agents the console supervises. Even for a caller confined to the interface, the two writable roots are whole home directories that between them hold the fleet registry, every instruction file, the permission settings file and the console's own source, so containment inside the allowlist is not containment of anything governance-relevant. The read side is wider still: the read route applies no sensitive-name filter, unlike the knowledge-base routes which do, so credential-shaped files under those roots are retrievable by any authenticated caller.
FAILURE MODE	A root added as writable for one convenience silently widens write, rename, upload and delete simultaneously, because scope is per root and not per route.
FAIL STATE	fail-closed
EMITTED SIGNAL	An audit line per mutating file operation and per download; list and read operations other than download emit nothing, so the largest surface, reading, is the least observed.
METRICS	MG-031 MG-059 MG-064
COST	low
FRICTION	low
AWS BINDING	EFS access points with an enforced root directory and posix identity, or S3 bucket policies with prefix conditions, so that the confinement is applied by the storage service rather than by the caller's code.
AZURE BINDING	Azure Files with identity-based access per share, or blob storage with container-scoped RBAC and stored access policies.
ON HOST	yes
EVIDENCE	EV-013 EV-017 EV-020
RESEARCH VALUE	An exemplar of a correct check placed where the governed party does not pass, which operationalises the difference between control quality and control coverage and gives structural enforcement coverage a concrete, arguable data point.

GR-L03-02 Namespaced per-session file root

INTENT	Give every agent session a distinct directory for its transcript, mailbox, metadata, token and ingested files, so that provisioning or deleting one session cannot collide with another or with an operational directory.
MECHANISM	Session paths are built under a dedicated root, one flat subdirectory per session keyed on the session name, with the name validated against a restrictive character class as the first statement of the path builder; a namespace resolver then realpaths the result and asserts containment. The root was deliberately chosen not to be the shared working home, because a session named after an operational directory would otherwise provision straight into it and the deletion cascade would destroy that directory.
ENFORCEMENT	conditional
BINDS	AC-014 AC-015 AC-023 AC-024 AC-028
ENABLES	II
PRECONDITION	That a session cannot reach the namespace except through the constructing code; at the filesystem layer on this host that is false.
BYPASS	The namespace resolver confines to the namespace root and not to the individual session, so even inside the interface one session's request can name another session's directory; per-session confinement exists only where it was written out explicitly, in the promotion route, where it is tight and correct. Outside the interface there is no confinement at all, because every session directory is owned by the same uid, so any agent reads, writes or deletes any peer's transcript, mailbox, metadata and bus token with ordinary tools. The design is nonetheless doing real work: a live session exists today whose name is one character away from an operational directory, and the namespacing is exactly what makes that harmless.
FAILURE MODE	A name that passes validation but matches a reserved entry at the namespace root would be treated as a session by the deletion cascade and as infrastructure by the retention reaper, with the two disagreeing about its lifetime.
FAIL STATE	fail-closed
EMITTED SIGNAL	Session provisioning and deletion events in the console audit log, plus the reaper's own audit entries; direct filesystem access by one agent to another session's directory emits nothing at all.
METRICS	MG-031 MG-040 MG-066
COST	low
FRICTION	low
AWS BINDING	One EFS access point per session, each pinned to its own root directory and posix identity, so the namespace is enforced by the kernel rather than by path construction.
AZURE BINDING	One Azure Files share or blob container per session with identity-scoped RBAC.
ON HOST	yes
EVIDENCE	EV-023
RESEARCH VALUE	It separates two things routinely conflated in isolation claims, namespace hygiene and access control, and does so within one system: the namespace here is well built and the access control is absent, which makes it a clean demonstration that isolation is a property of the enforcing layer.

GR-L03-03 Per-session filesystem isolation by namespace or container

INTENT	Make one agent's files unreachable by another agent, enforced by the kernel rather than by the code that constructs paths.
MECHANISM	Run each session in a container or mount namespace with only its own session directory bind-mounted, a distinct uid, and no visibility of peer directories, shared secrets or key material; anything genuinely shared is mounted read-only or brokered through a service.
ENFORCEMENT	structural
BINDS	AC-016 AC-022 AC-023 AC-024 AC-027 AC-053 AC-054
ENABLES	II
PRECONDITION	Per-agent identity and removal of passwordless root, since a root-capable process leaves a mount namespace it owns; also a settled decision about the shared context store, which every session currently depends on and which is keyed on the shared working directory.
BYPASS	With root retained the governed party enters another namespace or mounts whatever it likes, so the isolation is decorative until the identity layer changes; the ordering is not optional. There is also a design conflict here that cannot be bypassed but must be resolved, and it is the substantive cost of the control: the fleet deliberately starts every session in one working directory so that all agents share a single accumulated context store and a single transcript namespace, and per-session isolation severs that sharing by construction. Trading shared knowledge against isolation is the real decision, and it is not a configuration detail.
FAILURE MODE	A bind mount added for convenience reintroduces one shared writable path and quietly restores the full cross-session surface; a shared runtime socket offers an escape from the namespace it was meant to enforce.
FAIL STATE	fail-closed
EMITTED SIGNAL	Namespace or container creation and teardown events per session and denied access attempts at the mount boundary, produced by the runtime rather than by application code, so the signal does not depend on the agent's own instrumentation.
METRICS	MG-031 MG-038 MG-040 MG-062
COST	high
FRICTION	high
AWS BINDING	One ECS task or microVM per session with its own EFS access point, and Nitro Enclaves where the workload must be attested rather than merely separated.
AZURE BINDING	One Azure Container Instance or Container Apps revision per session with its own mounted file share and its own managed identity.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It would let the founding case's cross-session findings be re-measured with identity held constant, and it tests a specific and falsifiable hypothesis: that shared context, not the shared filesystem, is what actually drives measured agent effectiveness.

GR-L03-04 Read-only mounts for code the agent must not change

INTENT	Make the supervision machinery, the launch wrappers and the enforcement code physically unwritable by the processes they govern.
MECHANISM	Place wrapper scripts, hook scripts, unit files, settings and policy artefacts on a mount presented read-only inside the agent's namespace, with the writable copy held by a separate deployment identity; changes arrive by redeploying the mount rather than by editing in place.
ENFORCEMENT	structural
BINDS	AC-002 AC-021 AC-023 AC-025 AC-026 AC-032
ENABLES	II
PRECONDITION	Process isolation with a separate mount namespace, removal of root, and a deployment path that does not execute as the agent; without all three the mount is remounted writable in one command.
BYPASS	On this host one command as root remounts the path writable and the agent holds that command. The observed state makes the target concrete rather than theoretical: every hook script, wrapper, role file, roster and settings file is owned and writable by the agent, and the shared helper directory at the session namespace root is world-readable and agent-writable while holding the very scripts the hook mechanism executes on every tool call in every session. That is an unguarded self-modification path into the supervision layer. Made genuinely read-only from a separate deployment identity, this becomes one of the few controls in the catalogue the governed party cannot argue with rather than merely should not.
FAILURE MODE	An urgent fix cannot be applied in place during an incident, which generates sustained pressure to retain a writable escape hatch, and the hatch reinstates precisely the exposure the mount removed.
FAIL STATE	fail-closed
EMITTED SIGNAL	Write attempts refused with a read-only filesystem error, visible in the agent's own tool output and therefore directly countable, plus deployment events on the writable side recorded under the deployment identity.
METRICS	MG-031 MG-034 MG-035 MG-059
COST	moderate
FRICTION	high
AWS BINDING	EFS access points mounted read-only, or an immutable machine image layer with the artefact source held in S3 under Object Lock and applied by Systems Manager.
AZURE BINDING	Read-only blob or file share mounts backed by immutable storage policies, with artefacts deployed by a separate service principal.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It is the sharpest available test of the central claim, because it converts the proposition that an agent should not edit its own governance from an instruction into a kernel-level impossibility, and mandate widening initiated by the governed party becomes a count that should fall to zero.

GR-L03-05 Per-session storage quota

INTENT	Bound the disk a single session can consume so that one agent cannot exhaust the filesystem the whole fleet depends on.
MECHANISM	The ingestion route enforces a per-file ceiling inside the multipart handler, streaming and aborting mid-upload rather than buffering, caps a batch at twenty files, and after each file lands recomputes the session's total and deletes any file that pushes it past a per-session ceiling, returning an explicit rejection reason rather than truncating. The three governing values are unset in the environment, so the code defaults are in force.
ENFORCEMENT	conditional
BINDS	AC-023 AC-028 AC-053
ENABLES	II
PRECONDITION	That every write into the session directory passes through the ingestion route, which is not the case; a filesystem-level project quota would be needed for the control to be complete.
BYPASS	It is a route-level check and not a filesystem quota, so an agent writing into its own or a peer's directory with an ordinary file write or a shell redirect never encounters it. The accounting is also partial by design: it sums only the ingested-files subdirectory, leaving transcripts, mailboxes, the inbox and the shared extraction cache uncounted, and the largest single consumer observed on this host is an unbounded transcript. No project quota exists on the filesystem. What the control genuinely buys is worth keeping: an oversized or hostile upload through the operator interface fails fast, visibly, with a stated reason, and because the limit is applied to the stream a single large file cannot exhaust memory before it is rejected.
FAILURE MODE	Silent divergence between accounted usage and real usage, so the quota reports headroom while the volume fills from the categories it does not count.
FAIL STATE	fail-closed
EMITTED SIGNAL	An audit line per file recording acceptance or rejection with the reason and the extraction method used; nothing at all is emitted for the uncounted categories, which is exactly where the growth occurs.
METRICS	MG-031 MG-038 MG-059
COST	low
FRICTION	low
AWS BINDING	No native per-session quota exists for a shared instance filesystem and EFS access points do not bound size, so this is implemented in the workload, or session storage is separated onto its own volume or bucket with a size-bounded lifecycle policy.
AZURE BINDING	Azure Files per-share quotas give a native per-session ceiling where each session holds its own share; otherwise it is implemented in the workload.
ON HOST	partial
EVIDENCE	EV-023
RESEARCH VALUE	A compact demonstration that coverage rather than correctness determines effect: the enforced path is carefully built and governs a small minority of the bytes actually written to the namespace.

GR-L03-06 Immutable append-only evidence directory

INTENT	Make the record of what an agent did impossible for that agent to alter or remove after the fact.
MECHANISM	Write transcripts and audit records into a store that accepts appends and refuses modification and deletion for a defined retention period, with immutability enforced by the storage service rather than by file permissions, and forward records off the host as they are produced so that the local copy is a buffer and not the record.
ENFORCEMENT	structural
BINDS	AC-022 AC-053 AC-054 AC-055 AC-058
ENABLES	II
PRECONDITION	A separate identity holding the write permission, and a forwarding path that does not depend on the agent's cooperation to run.
BYPASS	Once a record is held off-host under a different principal with an immutability policy, the governed party genuinely cannot alter it, and this is one of the rare unambiguously structural cases in the catalogue. Two residual paths remain and must be stated. The agent still controls what enters the record, by acting outside the instrumented path, so immutability guarantees the integrity of what was captured and not the completeness of capture. And the local buffer before forwarding is writable, so a record can be altered in the window before commit; a synchronous commit closes that window and costs latency on every turn. On the present host none of this exists: transcripts carry no append-only or immutable filesystem attribute, are owned by the writing agent, and the backup that preserves pruned records lives on the same filesystem under the same uid.
FAILURE MODE	The store becomes unavailable and the pipeline either blocks the agent or discards records; a retention period set too short silently destroys the evidence the control was built to hold.
FAIL STATE	fail-closed
EMITTED SIGNAL	A commit acknowledgement per record carrying an object version, plus explicit denial events on attempted modification; the denial event is the strongest tamper signal available anywhere in this catalogue.
METRICS	MG-029 MG-064 MG-065 MG-066
COST	moderate
FRICTION	low
AWS BINDING	S3 with Object Lock in compliance mode and versioning, CloudWatch Logs in a log group the agent's role cannot delete, and CloudTrail with log file validation.
AZURE BINDING	Immutable blob storage with a time-based retention policy and legal hold, and Azure Monitor Logs with an immutability configuration.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The founding case's divergence measurement is only as good as the evidence it rests on, so this control is what makes such a measurement defensible rather than merely available, and it converts audit record immutability from an aspiration into a binary property of the storage.

GR-L03-07 Denial of write access to the agent's own instruction files

INTENT	Prevent an agent editing the mandate, role definition, roster and operating guidance that constitute its own instructions.
MECHANISM	Hold the instruction corpus under a separate identity, present it read-only inside the agent's namespace, require changes to arrive through a reviewed deployment, and verify at session start that the corpus actually loaded matches a recorded hash, refusing to start when it does not.
ENFORCEMENT	structural
BINDS	AC-019 AC-021 AC-023 AC-025
ENABLES	II
PRECONDITION	Read-only mounts and a deployment identity distinct from the agent; a hash check alone is defeated by an actor that can also update the recorded hash.
BYPASS	Today every instruction and instruction-adjacent file is owned and writable by the agent, including the operating guidance asserting standing change authority, the per-agent role files, the shared roster that carries the deployment approval chain, and the context index which is itself regenerated by an explicitly allow-listed, agent-invocable script. Three role files are symlinks into the same directory that holds the update plane's state and its halt switch, so one agent-writable directory holds both a mandate and the switch meant to stop it. Even with the corpus read-only a real path remains: instructions are consumed as model context, so an agent can change its own behaviour by writing to any other file that reaches context, which is why the control must name the whole context set and not only the files that look like instructions.
FAILURE MODE	A stale recorded hash after a legitimate change blocks every session start, and the pressure that creates leads to the check being downgraded to a warning, at which point it is telemetry rather than enforcement.
FAIL STATE	fail-closed
EMITTED SIGNAL	A session-start attestation recording the hash of the corpus actually loaded, plus refused write attempts on the corpus; at present the corpus can change with no audit event of any kind.
METRICS	MG-031 MG-034 MG-035 MG-050
COST	moderate
FRICTION	moderate
AWS BINDING	S3 with Object Lock as the source of truth, an EFS access point mounted read-only into the workload, deployment by a separate IAM role, and CloudTrail on the object writes.
AZURE BINDING	Immutable blob storage as the source, a read-only mount into the workload, deployment by a separate service principal, and Azure Monitor on the writes.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It operationalises the thesis directly: the same prose mandate held immutable on one arm and writable on the other, with mandate stability and self-initiated widening measured on each, is the experiment that distinguishes instruction from enforcement.

GR-L03-08 Denial of write access to the agent's own permission settings

INTENT	Prevent an agent widening the tool-layer permissions under which it is running.
MECHANISM	Place the deny list, allow list and default mode in an administrator-managed settings file owned by root and mounted read-only, positioned above the user and project scopes in the precedence order, and launch agents without the permission-skipping flag so that the file is in fact consulted.
ENFORCEMENT	structural
BINDS	AC-020 AC-025 AC-026 AC-029
ENABLES	II
PRECONDITION	The launch path must stop passing the permission-skipping flag and the default mode must not be the bypassing one; a protected settings file governs nothing while either holds.
BYPASS	On this host the control has two independent defeats and both are live. Every settings file is owned and writable by the agent, so it can add an allow rule, remove one of the twenty-three deny rules or change the default mode and continue in the same session; and, quite separately, the launch adapter passes the permission-skipping flag to every fleet agent, so the permission system is not consulted whatever any file says. No administrator-managed settings file exists anywhere on the host, so there is no layer above the writable ones at all. Making the file immutable therefore changes nothing measurable until the launch flag is removed, which makes this control strictly dependent on the process-layer entry that prohibits that flag.
FAILURE MODE	A managed layer that exists but is empty reads as enforcement while permitting everything, and precedence surprises let a lower scope silently win over the layer believed to be authoritative.
FAIL STATE	fail-closed
EMITTED SIGNAL	A refused write on the settings path, and a session-start record of the effective permission mode together with the settings files actually loaded; the latter does not exist today, which is why nominal and effective posture cannot currently be compared without manual inspection.
METRICS	MG-031 MG-035 MG-050 MG-059
COST	low
FRICTION	moderate
AWS BINDING	No native equivalent; implemented in the workload by holding the managed settings file on a root-owned read-only mount sourced from S3 with Object Lock and applied by Systems Manager.
AZURE BINDING	No native equivalent; implemented in the workload with an immutable blob source applied by a separate service principal and mounted read-only.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The pairing with the launch flag makes it a clean demonstration of defence-in-depth arithmetic, two controls each of which looks sufficient alone and which are jointly necessary, which is precisely what the defence in depth factor is meant to capture.

GR-L03-09 Soft delete with retention window for session files

INTENT	Make deletion of a session's files recoverable for a bounded period, so that an erroneous or hostile kill does not destroy state immediately.
MECHANISM	The session kill path renames the session directory into a trash area under the namespace root rather than removing it, and explicitly restamps the modification time because a rename preserves the original and would otherwise make the age of a trashed directory unreadable. A scheduled retention job runs daily under a service manager timer with a randomised delay and catch-up for missed runs; it moves ingested files past the retention period into trash and permanently purges anything trashed more than twenty-four hours earlier, with the hard delete gated behind a realpath guard. It writes its own audit entries under a system principal, and an audit write failure never aborts the run.
ENFORCEMENT	conditional
BINDS	AC-015 AC-024 AC-028 AC-054
ENABLES	II
PRECONDITION	That deletion occurs through the kill path; a direct removal by an agent has no soft-delete stage at all and no record.
BYPASS	The recovery window exists only for deletions routed through the console's kill cascade. An agent removing a directory with its own tools deletes it outright, and the same agent can empty the trash area early because it owns it. The window is also short, twenty-four hours from trashing, which is ample to undo an operator's mistake and far too short to serve as evidence retention, and the job's own audit trail is a file on the same filesystem under the same uid as the agent. What it genuinely delivers is the property its design comments target, that an erroneously or maliciously named session cannot cause an irreversible cascade, and the trash contents observed on the day of collection show the mechanism working exactly as designed.
FAILURE MODE	The timer being disabled leaves trash accumulating without bound and unnoticed; a failure of the restamping leaves entries whose apparent age is ancient, which the purge phase then removes immediately rather than after the window.
FAIL STATE	fail-closed
EMITTED SIGNAL	Retention job audit entries under a system principal covering expiry, purge, guard refusal and run completion, plus the trash directory itself as recoverable state; deletion performed directly on the filesystem by an agent emits nothing.
METRICS	MG-039 MG-043 MG-066
COST	low
FRICTION	low
AWS BINDING	S3 versioning with lifecycle rules and delete protection, or AWS Backup over the file system with a defined retention policy.
AZURE BINDING	Blob and container soft delete with a retention period, and Azure Backup for file shares.
ON HOST	yes
EVIDENCE	EV-023
RESEARCH VALUE	One of the few present controls on this host that is genuinely well engineered, down to a subtle correctness detail about rename and modification time, which makes it a valuable positive control when arguing that the weaknesses catalogued elsewhere are not the product of general carelessness.

GR-L03-10 Separation of the evidence store from the operational filesystem

INTENT	Ensure that the record of agent activity does not share a failure domain, an identity or a lifecycle with the systems the agents operate on.
MECHANISM	Ship transcripts, terminal captures and audit records off the host as they are produced, into a store held in a different account under a different identity, keeping the local copy only as a short buffer, and never allow a backup of the record to sit under the same uid on the same filesystem as the record.
ENFORCEMENT	structural
BINDS	AC-022 AC-053 AC-054 AC-055 AC-058
ENABLES	II
PRECONDITION	A separate identity holding the write permission and a network path that survives egress restriction; also an explicit decision about what the local buffer may hold and for how long.
BYPASS	An agent cannot alter what has already been shipped, but it keeps two levers. It controls the local buffer, so anything not yet forwarded remains editable, and it controls whether an action is instrumented at all, so it can act outside the captured path entirely. On the present host the separation is absent in an instructive way rather than simply missing: the frequent backup of the transcript tree deliberately omits deletion so that pruned records survive, which is thoughtful retention design, but it lands on the same host, the same filesystem and the same uid, so it is a copy rather than an independent record. The separate terminal captures are raw byte streams with no shared record identity with the structured transcripts, so even where both survive the two cannot be correlated.
FAILURE MODE	A forwarding backlog during a network outage silently discards records once the buffer wraps, and a store placed in the same account inherits the same administrative identity, leaving the separation nominal while the architecture diagram shows it as real.
FAIL STATE	fail-closed
EMITTED SIGNAL	Per-record delivery acknowledgements with sequence numbers, and a gap or backlog alarm when delivery falls behind; the gap alarm is what distinguishes a quiet agent from a broken evidence pipeline, which no present signal can do.
METRICS	MG-048 MG-064 MG-065 MG-066
COST	moderate
FRICTION	low
AWS BINDING	CloudWatch Logs or Kinesis Data Firehose delivering to S3 in a separate account with Object Lock, a cross-account role holding write-only permission, and an organisation-level CloudTrail.
AZURE BINDING	Azure Monitor diagnostic settings delivering to a Log Analytics workspace in a separate subscription, or immutable blob storage with cross-tenant write-only access.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It makes episode reconstructability independent of the agent's cooperation, which is the precondition for treating any divergence measurement as evidence rather than as testimony taken from the party under study.

GR-L03-11 **Working tree isolation per agent**

INTENT	Give each agent its own checkout so that concurrent agents cannot sweep each other's uncommitted changes into a commit or race on the same branch.
MECHANISM	Provision a git worktree or separate clone per agent session, bound to a branch named for that session, with the shared clone presented read-only; changes return to the mainline through the normal review path rather than through direct writes into a common tree.
ENFORCEMENT	conditional
BINDS	AC-002 AC-003 AC-004 AC-005 AC-008
ENABLES	II
PRECONDITION	Per-session filesystem isolation, or at minimum a read-only mount on the shared clone; otherwise an agent simply changes directory into the shared tree and the separation is a habit.
BYPASS	With a shared uid and no mount separation an agent reaches any tree by path, so the isolation is convention rather than boundary. The concrete hazard it addresses is already recognised on this host: a subagent or peer working in the same tree has its uncommitted work swept into another agent's commit, which is a correctness failure rather than a security one and is invisible in the resulting history. Making the shared clone read-only converts convention into boundary at the cost of an explicit merge step. It is worth recording why this sits at the weaker class rather than higher: on this host committing and pushing are treated as version control only and are not a deployment gate, so the blast radius of a bad commit is bounded by the fact that deployment happens separately and on the box.
FAILURE MODE	Worktree proliferation exhausts disk, and a stale worktree still bound to a deleted branch blocks later operations with errors that are hard to attribute to their cause.
FAIL STATE	fail-open
EMITTED SIGNAL	One branch and one commit trail per agent, making authorship attributable in the version control history itself; today commits from several agents in one tree are indistinguishable except by message convention.
METRICS	MG-007 MG-009 MG-067
COST	low
FRICTION	moderate
AWS BINDING	No native equivalent; implemented in the workload with per-session storage, optionally one EFS access point per session so that the tree is also isolated by the kernel.
AZURE BINDING	No native equivalent; implemented in the workload with per-session file shares or ephemeral container storage.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	It isolates a failure mode that is pure concurrency rather than security, which is valuable because it shows that agent isolation pays for itself in rework and change failure terms before any governance argument is made at all.

2.4 L4. Processo e execução

Esta camada governa como os processos de sistema operacional de um agente são iniciados, limitados e encerrados. O controle presente mais forte de todo o catálogo, em termos de qualidade de implementação, está aqui: a criação de processos usa argumentos em array em vez de uma string de shell em essencialmente todo ponto de chamada, o que remove, de modo structural, a injeção de metacaracteres de shell como vetor nos caminhos que a utilizam **EV-017**. A entrada presente mais fraca da camada está bem ao lado: a supressão de verificação de permissões vem habilitada por padrão para os agentes da frota, fixada no código sem opção de desativação e reafirmada uma segunda vez como um bypass padrão em todo o host, presente e classificada `none` **EV-032**. O único portão de permissão em vigor por trás desse padrão é uma lista de negação de comandos com cerca de vinte e quatro padrões de shell, conditional em vez de structural porque não nomeia nenhuma CLI de nuvem,

nenhum shell remoto, nenhum gerenciador de processos, nenhum comando de escalonamento de privilégio e nenhum caminho de escrita de arquivo, e porque o arquivo de configurações que a contém é gravável pela mesma conta que ela deveria governar EV-032.

O encerramento automático por detecção de travamento (wedge) é parcial: existe para nove das catorze sessões ativas, porque o escopo dos observadores (watchers) é um array fixo no código com nomes de agentes, em vez de uma enumeração ao vivo do que está em execução EV-021 EV-028. Um controle que meio-existe dessa forma não é uma versão menor do controle completo; é uma coisa diferente: um agente fora do array não tem detecção de travamento alguma, e o catálogo registra a entrada como `partial` em vez de `present` exatamente por esse motivo.

Camada 4. Processo e execução.

GR-L04-01 Agent name validation by regular expression

INTENT	Constrain the agent and session identifier character set so that a name can never be read as a shell metacharacter, a terminal multiplexer target separator or a path traversal segment.
MECHANISM	Two identical regular expressions, one in the console tmux control layer and one in its fleet registry module, permitting only letters, digits, dot, underscore and hyphen up to 80 characters. The assertion is the first statement of every exported tmux primitive (session create, has-session, capture, send-keys, paste, scroll, pipe attach, kill, display) and of every fleet registry helper, so it is applied at the primitive layer and every console caller inherits it rather than each caller remembering. It rejects whitespace, slash, colon, quotes, backtick, dollar, semicolon, pipe, ampersand, parentheses, brackets, backslash and all non-ASCII; rejecting the colon is what prevents session and window target confusion in send-keys and kill.
ENFORCEMENT	conditional
BINDS	AC-014 AC-015 AC-016 AC-017 AC-019
ENABLES	II
PRECONDITION	Every path that creates or addresses a session must route through the console tmux primitive layer for the check to be complete.
BYPASS	Three concrete paths route around it. First, a raw terminal multiplexer new-session command from any shell performs no validation and produces no audit record; the live multiplexer server on this host was itself created that way and every agent is a child of it. Second, the fleet wrapper takes its agent name as an unvalidated positional argument and immediately builds filesystem paths and multiplexer targets from it, so anyone who can run the script or edit a unit file supplies whatever name they like; the value is validated one hop upstream at registry insert time and nowhere else. Third, the registry shell writer performs existence checks only (agent exists, group exists, tool wired, model valid) and never re-checks the character set, so a direct command line call inserts a badly named agent. Underneath all three, the account that runs every agent holds unrestricted passwordless root, so no file mode prevents any of them. What the control still buys is real: it removes name-derived injection and target confusion from every console-mediated path, which is the only path an operator browser can reach.
FAILURE MODE	If the expression is widened or the assertion dropped from one primitive, a name containing a colon reaches the multiplexer as a session and window target, and send-keys, capture and kill then act on a different pane than intended, with no error.
FAIL STATE	fail-closed
EMITTED SIGNAL	A rejected name produces an HTTP 400 with a fixed message and, on the fleet creation path, a console audit line; both are observable in the console audit log and backend log. Names created through the raw multiplexer path emit nothing at all, which is the coverage gap expressed as a signal gap.
METRICS	MG-032 MG-033 MG-059
COST	low
FRICTION	low
AWS BINDING	No native equivalent; implemented in the workload. The nearest analogue is a Systems Manager Automation or Run Command document parameter with an allowedPattern, which validates the parameter before the document executes, and IAM resource naming condition keys for the creation half.
AZURE BINDING	No native equivalent; implemented in the workload. Azure Policy with a name pattern deny effect at resource creation is the closest control-plane analogue, and Azure Automation runbook parameter validation the closest execution-plane one.
ON HOST	yes
EVIDENCE	EV-017 EV-020 EV-022
RESEARCH VALUE	A clean case of a control applied at the correct layer whose coverage over the action class is nonetheless partial because six other creation paths exist. It lets coverage within a code path be separated experimentally from coverage over an action class: enabling and disabling the raw multiplexer path changes the unclassified action share without changing one line of the control itself.

GR-L04-02 Process spawn with array arguments rather than a shell string

INTENT	Prevent any value that a request or an agent supplies from being parsed as shell syntax when the console creates an operating system process.
MECHANISM	An exhaustive and verified discipline in the console backend. A shell option is never set anywhere in the source; the only process creation functions imported from the Node child process module are execFile and execFileSync, across 13 and 3 modules respectively, and the shell-invoking variants are never imported at all. The only other process creation is a pseudoterminal spawn used in three places, also with argument arrays. All 34 call sites pass an argument vector, never a command line as argument zero. Where a shell is genuinely required the script text is one array element handed to a login shell or to a remote secure shell daemon. The search call sites add fixed-string matching, one explicit pattern flag per token, a double-hyphen terminator and a resolved real path for the search root, which is the strongest input handling in the codebase.
ENFORCEMENT	conditional
BINDS	AC-012 AC-014 AC-023 AC-029 AC-043
ENABLES	II
PRECONDITION	A single reviewed spawn helper, and a standing rule that no module imports the shell-invoking process creation functions; today the rule is upheld by convention and has no lint rule or continuous integration check behind it.
BYPASS	The discipline is not defeated so much as circumvented by its own design. Eight sites hand a constructed script string to a login shell, to the multiplexer (which runs pane and pipe commands through a shell of its own) or to a remote login shell, so the array boundary is perfect locally and absent one level down. In the update plane a single regular expression over package and version tokens is the only thing between strings returned by a possibly compromised remote host and a privileged package install on nine machines. The deeper point is that the discipline is irrelevant to what the console is actually for: the terminal WebSocket writes raw browser bytes into a live agent pseudoterminal, and the chat and message bus interrupt modes paste operator or peer text into a running pane. Text delivered to an autonomous agent is not an argument to be escaped, it is an instruction to be obeyed, and only extracted attachment content is wrapped in a treat-as-data envelope; operator chat, bus messages between agents and broadcast text carry none. What it still buys is that the console own 34 spawn sites are genuinely uninteresting to an attacker, which is a rare property and worth crediting.
FAILURE MODE	One new call site that sets the shell option or concatenates a command line reintroduces argument injection silently, because nothing tests for it; the property is currently maintained by reviewer memory.
FAIL STATE	fail-open
EMITTED SIGNAL	None today. No spawn site emits a structured event, so an array spawn and a shell spawn are indistinguishable from outside the process. The signal it would have to emit is a per-spawn record naming the binary, the argument vector and whether any shell interpreter was in the resulting chain, written to a stream the spawning account cannot edit.
METRICS	MG-032 MG-033 MG-059
COST	low
FRICTION	low
AWS BINDING	No native equivalent at the process layer; the architectural replacement is Systems Manager Run Command documents with typed parameters and allowedPattern, and Lambda invoked with a structured event rather than a command string, so no shell parses caller data at any level.
AZURE BINDING	No native equivalent; Azure Automation runbooks with typed parameters, and Container Apps jobs invoked with an argument array rather than a shell entrypoint.
ON HOST	yes
EVIDENCE	EV-017 EV-020
RESEARCH VALUE	The clearest case in the estate of a control that is genuinely well executed and yet governs almost none of the risk, because the risk sits one layer below it and one layer above it at the same time. That makes it the ideal negative control in an experiment: toggling it should not move the boundary crossing success rate at all, which tests

> whether the measurement instrument can distinguish application security from agent governance.

GR-L04-03 Seccomp or mandatory access control profile on agent processes

INTENT	Reduce the kernel-level capability surface of an agent process so that whole classes of system call and whole regions of the filesystem are unreachable regardless of what the process decides to do.
MECHANISM	Would be: run each agent under a container runtime or a systemd unit carrying a seccomp profile denying mount, ptrace, kexec, bpf, module load and raw device system calls, together with an AppArmor or SELinux profile confining the process to its own working tree, its own session storage and the specific tool binaries its role needs. The instruction corpus, the permission settings file and the update plane kill switch directory are mounted read-only into the confined view. Expressed as unit directives that is NoNewPrivileges, ProtectSystem set to strict, ProtectHome as a tmpfs with explicit bind paths, PrivateTmp, an empty CapabilityBoundingSet and a SystemCallFilter restricted to the service set.
ENFORCEMENT	structural
BINDS	AC-022 AC-023 AC-024 AC-025 AC-026 AC-027 AC-029 AC-032
ENABLES	II
PRECONDITION	Per-agent process boundaries must exist. Today the nine agent units deliberately set the kill mode to process precisely because the multiplexer server is shared across every session, so the control group is not the agent and there is no per-agent process tree to confine.
BYPASS	A profile attached to the agent unit does not follow a process the agent starts outside that unit, and on this host the multiplexer server is a long-lived process with init as its parent that no unit owns; confining the agent while leaving the multiplexer unconfined confines nothing. Where the account keeps passwordless root the agent edits the unit or a drop-in and restarts itself unconfined, so this control is only meaningful once the root grant is removed, which makes the identity layer its true precondition. Note that AppArmor is already enabled on this host at the service level and no agent unit carries a profile at all, so adoption is a configuration act rather than a platform change, and the absence is a choice.
FAILURE MODE	An over-tight profile terminates legitimate tool use at an unpredictable point, typically mid-task, leaving partial work and no clean failure record; the estate current wedge detector would read the result as a stall and type a sentence into a session whose working tree is already inconsistent. An under-tight profile is theatre that raises the measured coverage without changing reachability.
FAIL STATE	fail-closed
EMITTED SIGNAL	Kernel audit records for each denied system call and each denied file access, through the audit subsystem and the mandatory access control denial log. This is one of the few controls in the catalogue whose signal is independent of the agent by construction, because the kernel writes it and the agent cannot suppress it without defeating the control itself.
METRICS	MG-031 MG-033 MG-040 MG-042
COST	moderate
FRICTION	moderate
AWS BINDING	ECS or Fargate task definitions carrying a custom seccomp profile through linuxParameters with a distinct task role per agent, EKS pod security standards with a restricted baseline, and Nitro Enclaves where the isolation must survive compromise of the host itself.
AZURE BINDING	Container Apps or AKS with a seccomp profile and pod security admission, with Azure Policy set to deny any workload deployed without the profile so the confinement cannot be omitted by a deployment.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The cleanest available test of the central thesis claim, because it converts a prohibition currently expressed only as prose into a kernel refusal at the same point in the same run. The experiment is to execute identical briefs with and without the profile: a structural control should drive boundary crossing success rate to zero for the confined classes, and

	any drop in task completion rate is a direct measurement of the real friction cost of structure rather than an assertion about it.

GR-L04-04 Control group limits on agent processes

INTENT	Bound the compute, memory, task count and input output an agent may consume so that one agent cannot starve the host, its supervisor or its peers.
MECHANISM	Would be: a per-agent systemd slice carrying MemoryHigh and MemoryMax, CPUQuota, TasksMax and an input output weight, plus a process identifier ceiling to cap fork depth, so that a runaway loop is first throttled and then terminated inside its own slice rather than taking the shared multiplexer server and every sibling agent with it.
ENFORCEMENT	structural
BINDS	AC-008 AC-011 AC-012 AC-014
ENABLES	I
PRECONDITION	One control group per agent, which requires each agent to own its process tree. The present units set the kill mode to process for the express reason that the multiplexer server is shared, so the cgroup boundary and the agent boundary are not the same object today.
BYPASS	There is nothing to defeat in the limit itself, since a control group ceiling is enforced by the kernel and no user space decision participates. It is routed around entirely by placement: a process the agent starts through the shared multiplexer server, or through one of the nine process-manager-supervised daemons, is not in the agent slice, and with passwordless root the agent writes a drop-in and reloads it. The more important limitation is categorical rather than technical: a control group bounds a resource footprint, not an authority, so an agent confined to one processor and half a gigabyte can still delete a production database. It belongs in the catalogue as an availability control, not a governance one, and should be scored as such.
FAILURE MODE	A limit set below the genuine working set of a build or a model run causes silent out-of-memory termination mid-task; because the estate measures liveness by session existence and progress by a regular expression over rendered terminal text, that termination is indistinguishable from a stall and the recovery path nudges a dead session.
FAIL STATE	fail-closed
EMITTED SIGNAL	Control group accounting and out-of-memory kill records in the system journal, plus memory pressure events per slice, all emitted by the supervisor rather than by the agent.
METRICS	MG-002 MG-038 MG-068
COST	low
FRICTION	moderate
AWS BINDING	ECS or Fargate task level processor and memory reservations, Lambda with reserved and provisioned concurrency to cap parallel execution, and instance level systemd slices for workloads that stay on EC2.
AZURE BINDING	Container Apps per-revision processor and memory limits with scale rules, AKS resource requests and limits enforced by a LimitRange, and Azure Policy requiring both.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Isolates the resource dimension of blast radius from the authority dimension, which the register otherwise conflates. Running an identical brief under a tight and a loose slice tests whether contention rather than authority is what degrades a multi-agent fleet, a proposition the estate own restart counts on its search service and console suggest but cannot establish.

GR-L04-05 Tool-layer command deny list

INTENT	Stop a small enumerated set of catastrophic shell commands from being issued through the model tool layer.
MECHANISM	Twenty three string glob rules in the harness permission settings file, each of the form Bash followed by a command pattern, matched against the command string before the shell tool runs it. They cover recursive deletion of root or home in nine spellings including two Windows shapes, force push to the two conventional default branch names in five spellings, raw block device writes, filesystem creation, host shutdown, host reboot, one fork bomb shape, and a download piped into a shell in four spellings. The same file contains no allow array and no ask array, and sets the default permission mode to bypass, so anything not matched by a deny rule executes with no prompt; a sibling key additionally suppresses the warning that would otherwise announce that posture.
ENFORCEMENT	conditional
BINDS	AC-006 AC-024 AC-029 AC-031 AC-036
ENABLES	III
PRECONDITION	The settings file must be loaded by the harness at session start and must not be writable by the process it governs; the second half is false here, and a filesystem-wide search found no administrator-owned managed settings file anywhere on the host.
BYPASS	A deny list enumerates the bad and therefore fails open on everything unenumerated. These rules match command spelling, not capability: recursive deletion of the home directory contents by absolute path is not listed, the same deletion under sudo is not listed, and every one of the 23 patterns is evaded by a shell variable, a subshell, a here-document, a script file written a moment earlier, or the same system call issued from a Python or Node one-liner. Independently, the file is owned by the account every agent runs as and is writable by it, with no managed tier above it, so an agent removes a rule and continues in the same session with no audit event. Decisively, the fleet launch catalogue passes the permission-skipping flag on the command line for all nine registered agents and the default mode is already bypass, so this list is the entire remaining tool-layer surface on the box. What it still buys, and it is not nothing, is that it removes the single most common accidental self-destruction, a recursive delete typed at the wrong root, from the space of easy mistakes; that is an availability benefit, not a security one, and the register should score it that way.
FAILURE MODE	A rule with a slightly wrong glob matches nothing and looks identical to a rule that is simply never triggered, because there is no rule hit counter, no test and no report; the list can therefore rot without any observable change.
FAIL STATE	fail-open
EMITTED SIGNAL	A denial surfaces to the agent as a tool error inside its own session transcript and nowhere else. There is no counter, no aggregate and no operator-facing record, so no party other than the agent learns that a boundary was touched. The signal it would have to emit is a per-denial event carrying the matched rule identifier to a store the agent cannot write.
METRICS	MG-032 MG-033 MG-059 MG-060
COST	low
FRICTION	low
AWS BINDING	No native equivalent at command level; the architectural replacement acts on capability rather than spelling, namely an IAM or Service Control Policy deny on the API the command would ultimately call, or Session Manager restricted to a fixed document set.
AZURE BINDING	No native equivalent at command level; Azure Policy deny effects on the corresponding control plane operation, or a restricted Azure Automation runbook catalogue, both of which bind the operation rather than the string that requests it.
ON HOST	yes
EVIDENCE	EV-020 EV-028 EV-032
RESEARCH VALUE	The catalogue canonical example of a control that is technically a runtime check and governs nothing. Paired with the allowlist entry it gives a single-variable contrast between deny and permit semantics at the identical enforcement point: the prediction is that boundary crossing success rate stays high under the deny list and collapses under

> the allowlist, while task completion rate falls only under the allowlist, so the pair prices the completeness of enumeration directly.

GR-L04-06 Tool-layer command allowlist

INTENT	Make the set of commands an agent may execute finite, enumerated and reviewed, so that an unrecognised command is refused rather than permitted.
MECHANISM	Would be: replace the deny array with an allow array and an ask array under a default mode of deny, so that every shell invocation is matched against enumerated command shapes appropriate to the role, with anything unmatched refused or escalated to a human decision. Entries must be parameterised commands rather than wildcards. The estate own existing allow file is the anti-pattern to design against: 35 rules including unbounded wildcards for a shell, an interpreter, a version control client, a remote shell client and a file synchronisation client, each of which permits arbitrary execution on its own, and one rule that embeds a database connection string with its password. The allowlist must live in an administrator-owned managed settings file outside the agent account write scope.
ENFORCEMENT	conditional
BINDS	AC-010 AC-011 AC-024 AC-029 AC-031 AC-036 AC-042 AC-043
ENABLES	II
PRECONDITION	A managed settings tier owned by an identity the agent does not hold. None exists: a filesystem-wide search found no managed settings file, and all eight settings files present are owned and writable by the agent account.
BYPASS	An allowlist enumerates the good and therefore fails closed on anything unenumerated, which is the exact inversion of the deny list and the whole reason to build it. Its defeats are different in kind rather than in degree. One over-broad entry restores full execution, because permitting any invocation of a shell, an interpreter or a remote shell client is equivalent to permitting everything; the estate existing allow file already contains four such entries. Matching remains string based, so a permitted verb can carry an arbitrary payload, for instance a permitted interpreter executing a script the agent wrote in the previous tool call, which means the allowlist bounds the verb and not the effect. On this host it would additionally be defeated by the same file ownership problem as the deny list until the managed tier exists, and by the permission-skipping flag, which disables the permission system entirely regardless of what any settings file says.
FAILURE MODE	An incomplete allowlist blocks legitimate work at unpredictable points, and the block presents to the agent as a tool error it will attempt to route around, which is empirically the condition that produces the most inventive bypasses; the control therefore generates the behaviour it exists to prevent unless the refusal is also escalated to a human.
FAIL STATE	fail-closed
EMITTED SIGNAL	A refusal event per unmatched command carrying the attempted command shape and the role that attempted it. That stream is directly the boundary crossing attempt rate, and it must be written to an operator-visible store rather than only returned to the agent, otherwise the control produces data only the governed party can see.
METRICS	MG-031 MG-032 MG-036 MG-060
COST	moderate
FRICTION	high
AWS BINDING	IAM identity policies with explicit allow and implicit deny applied to the API the command would call, Session Manager restricted to an approved document set, and Service Control Policies as the account-wide floor beneath both.
AZURE BINDING	Azure RBAC custom roles with enumerated actions and an implicit deny default, a deny-by-default Azure Policy initiative, and a restricted Azure Automation runbook catalogue.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The direct experimental counterpart to the deny list. Because both act at the same point in the same tool layer, swapping one for the other is a genuine single-variable manipulation, which makes it the cleanest available test of whether enumeration direction rather than enforcement location determines boundary crossing success and the fail closed ratio.

GR-L04-07 Wall clock timeout per agent turn

INTENT	Bound the elapsed time a single agent turn may run, so that an unattended agent cannot act indefinitely without reaching a decision point.
MECHANISM	Would be: a supervisor-side deadline per turn. The launcher records a turn start when a prompt is delivered, and at expiry sends an interrupt to the session, writes a turn timeout record, and requires either an explicit operator resume or an automatic downgrade to a read-only posture. It runs outside the agent process so the agent cannot extend its own deadline. The estate already contains the exact shape in dormant form: a general purpose loop harness with a default iteration cap of ten, a total timeout of 240 minutes, exponential backoff and a dead-man file that stops the loop when removed. No job is defined for it and no live session is bounded by anything.
ENFORCEMENT	conditional
BINDS	AC-014 AC-019 AC-036 AC-043
ENABLES	II
PRECONDITION	A turn boundary observable from outside the agent. Today the only external turn signal is a transcript hook event, and liveness is measured by session existence alone, so the supervisor cannot presently tell when a turn began.
BYPASS	A deadline terminates a turn and does not undo what the turn already did, so it bounds duration rather than effect; an agent that reaches a destructive action in the first ten seconds is entirely unaffected by a thirty minute ceiling. It is routed around by detachment: work started with a background process, a new process-manager entry or a new unit outlives the turn and is not interrupted, and the account can do all three. Where the timer runs on the same host under the same account, the agent can also stop the timer. What it still buys is the precondition for everything temporal: it converts an indefinite unattended run into a bounded one, which is what any re-authorisation regime requires in order to have something to re-authorise.
FAILURE MODE	A deadline that fires during a long legitimate operation such as a build or a package install leaves partial state with no rollback, and the estate current recovery path, a cron job typing a sentence into the pane, will resume a session whose working tree is already inconsistent.
FAIL STATE	fail-closed
EMITTED SIGNAL	A turn timeout event naming the turn identifier, the elapsed time and the action classes attempted within the turn, on the supervisor log rather than in the session transcript.
METRICS	MG-004 MG-044 MG-046 MG-069
COST	low
FRICTION	moderate
AWS BINDING	Step Functions state and heartbeat timeouts wrapping each agent invocation, Lambda function timeout for headless runs, ECS task stop timeout, and EventBridge Scheduler to arm the deadline outside the workload.
AZURE BINDING	Durable Functions orchestration timeouts, Container Apps job replica timeout, and Logic Apps action timeouts for the supervisory half.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	There is no idle timeout and no maximum session lifetime anywhere on this host, and one live session produced no terminal output for seven consecutive days while remaining listed as live and desired. A turn deadline creates the first bounded unit of unattended work in the estate, which is the unit every temporal metric needs; it is what makes latency margin computable at all.

GR-L04-08 Automatic termination on wedge detection

INTENT	Ensure that an agent which has stopped making progress is stopped and replaced, rather than left occupying a supervised slot while appearing alive.
MECHANISM	Present in a degraded form. Three cron watchers derive a stall signal from a regular expression over the last six rendered lines of the pane combined with the multiplexer window activity age, at thresholds of 180 seconds for the stall watcher, 150 seconds for the heartbeat watcher and 300 seconds for the human-facing watcher. On detection none of them terminates: two type an English sentence into the pane and press Enter, and the third only messages the operator. A separate branch auto-answers the model tool own resume chooser by sending the affirmative option. The only genuine termination is the daily scheduled reload, in which a unit restart runs a graceful stop that pastes a save-your-state prompt, polls a hash of the pane tail for up to 35 seconds and then kills the session; the units carry restart-always with a ten second delay and a kill mode of process so that stopping one agent cannot tear down the shared multiplexer server.
ENFORCEMENT	conditional
BINDS	AC-012 AC-014 AC-015 AC-055
ENABLES	III
PRECONDITION	A progress signal distinguishable from a liveness signal. None exists: every observation channel on the host reports existence or output, and none reports progress, so wedge detection is necessarily inferential.
BYPASS	Nothing has to defeat this control because it rarely acts, and its three weaknesses are independent. Coverage: 8 of the 14 live sessions have no stall detection of any kind, including three registry-backed fleet agents created on the day of collection and every ad-hoc session, and the target lists are hardcoded arrays in files the supervised agents can write. Signal: the busy expression matches the model tool spinner wording, so an upstream change to that wording disables stall detection fleet-wide with no test and no alert; two production failures of exactly that fragility are recorded in the scripts own comments, one of which left a session frozen at an interactive chooser for about an hour and three quarters because that chooser displays no busy marker at all. Response: even on detection the mechanism types a sentence rather than terminating, so a genuinely wedged process stays wedged and the escalation is a message. Finally, on the fleet launch path the model process is a child of a login shell, so if the model exits the pane survives, the multiplexer still reports the session, the registry still reports the agent live, and neither the wrapper nor the supervisor ever restarts it.
FAILURE MODE	The detector reads a genuinely idle session as busy through a static completed-turn footer left in scrollback and therefore never fires, or fires indefinitely against a session that cannot respond, producing repeated nudges and a stream of escalations that trains the operator to ignore the channel.
FAIL STATE	fail-open
EMITTED SIGNAL	Per-agent state files and a watchdog log recording detections and nudges, plus an escalation message to the operator after a streak threshold. It records the detector opinion rather than the agent state, and it is written by a process running under the same account as the agents it polices.
METRICS	MG-013 MG-017 MG-044 MG-068
COST	moderate
FRICTION	low
AWS BINDING	ECS or Fargate container health checks with essential-task replacement, Auto Scaling health checks that replace rather than notify, and Step Functions heartbeat timeouts with a catch branch that terminates the execution.
AZURE BINDING	Container Apps liveness and readiness probes with replica replacement, and Azure Monitor action groups triggering an Automation runbook that stops and recreates the instance.
ON HOST	partial
EVIDENCE	EV-005 EV-020 EV-027 EV-028
RESEARCH VALUE	The estate best natural experiment on signal quality, because detection here is a regular expression over a rendered terminal image, that is, a proxy for a proxy. Comparing its detection latency against a real progress signal measures how much of the wedged

> session detection latency is attributable to the instrument rather than to the fault, which is the same question that dogs every self-report-based oversight channel in the case.

GR-L04-09 Prohibition of the permission-skipping flag in defined contexts

INTENT	Make it impossible to launch an agent with the model tool permission bypass in contexts where prior approval is required, so that the interactive permission prompt survives as a decision point.
MECHANISM	Would be, and it needs two parts because the flag is asserted in five places in three languages. First, an administrator-owned managed settings file outside the agent account write scope that pins the permission mode and refuses a command line override. Second, launcher-side policy: the fleet tool catalogue, the console launch builder, the headless run endpoint and the browser terminal server each resolve their flag set from one signed policy document keyed by context, with a refusal at startup if the resolved flag set does not match policy. Contexts would distinguish development-only work, where bypass may be acceptable, from any context that can reach a production surface, where it may not.
ENFORCEMENT	structural
BINDS	AC-014 AC-020 AC-021 AC-026 AC-059
ENABLES	II
PRECONDITION	A settings tier the agent account cannot write, and a single resolution point for launch flags. Today the flag is hardcoded independently in a TypeScript constant, a JSON tool catalogue, a shell fallback and two JavaScript argument arrays, and there is no environment variable, configuration file or interface toggle anywhere that turns it off.
BYPASS	The prohibition binds launchers, not agents, so an agent that can start a process can start a model session itself with whatever flags it chooses; on this host that is one raw multiplexer command away, and the multiplexer server itself was created by exactly such a command. It is further defeated by editing any of the five hardcoded sites, all of which are agent-writable, or by writing the flag into the stored desired-session record, which the console replays verbatim at startup; the live evidence already shows one running process carrying a flag that no code path emits, so the stored record describes intent rather than reality. Making the prohibition genuinely structural therefore requires the model binary itself to refuse the flag when a managed policy forbids it, which is a property of the tool rather than of this estate, and that dependency should be stated plainly rather than assumed away.
FAILURE MODE	A partial rollout leaves one launch path still passing the flag and is indistinguishable from success until an agent uses that path; the estate already exhibits the shape, since even the degraded fallback in the fleet wrapper passes the flag when the tool catalogue lookup fails.
FAIL STATE	fail-closed
EMITTED SIGNAL	A launch-time record of the resolved flag set per session, and a refusal event when a launcher requests a forbidden flag. The resolved posture should also appear in the session listing, so that an operator can read the permission posture of every live session without inspecting process command lines.
METRICS	MG-030 MG-031 MG-050 MG-058
COST	low
FRICTION	high
AWS BINDING	No native equivalent for a tool flag; the architectural analogue is a permission boundary or Service Control Policy that makes the elevated path unusable even when requested, with Config rules detecting task definitions that carry the forbidden argument.
AZURE BINDING	Azure Policy with a deny effect on container specifications carrying the forbidden argument, enforced at admission, plus Machine Configuration to pin the host-level settings file.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The single highest-leverage variable in the catalogue, because the flag it prohibits is what converts every tool call from a decision point into an action. An experiment varying only this flag, with brief and model held constant, isolates the causal contribution of the human confirmation prompt to boundary crossing success rate and to task completion rate, and therefore prices human-in-the-loop supervision directly rather than by assertion.

GR-L04-10 Permission skipping enabled by default for fleet agents

INTENT	Allow long-running unattended work to proceed without a human present to answer the model tool interactive permission prompt, so that a scheduled or watcher-triggered agent completes a task overnight instead of stalling at a confirmation.
MECHANISM	The bypass is asserted redundantly at every layer, which is why it is effectively irreversible in practice. The fleet tool catalogue carries the permission-skipping flag as a standing extra argument for the wired tool, and the wrapper hardcoded fallback passes it too, so even the degraded path bypasses. The console launch builder holds it as a TypeScript constant, with equivalents for the three other supported tools, one of which also disables that tool sandbox, and the code states the intent outright as fully autonomous from the first keystroke with no human-in-the-loop confirmation ever required. The headless run endpoint and the browser terminal server each hardcode it in a JavaScript argument array. Independently the harness settings file sets the default permission mode to bypass and suppresses the warning banner that would announce it. All 14 live processes were observed carrying the flag on their command line.
ENFORCEMENT	none
BINDS	AC-016 AC-023 AC-024 AC-029 AC-031 AC-036 AC-039 AC-043
ENABLES	I
PRECONDITION	An operating decision that unattended completion matters more than per-action confirmation, together with the absence of any managed policy tier that could refuse the flag.
BYPASS	There is nothing to bypass, because the control is absent by construction; this row exists so that the catalogue can represent the deliberate removal of a control as a first-class object rather than as a silence in the record. The useful question is inverted: what would have to be true for the prompt to return. Four things. The flag would have to be removed from five hardcoded sites in three languages. The settings default mode would have to change. A managed tier the agent account cannot write would have to exist, and none does anywhere on the host. And something other than a cron job would have to answer the prompts, because a separate automation already presses the affirmative option in the model tool own resume chooser on the agent behalf. Absent all four, every action every agent takes on this host is unprompted by design, and the tool layer only remaining check is a 23 pattern deny list.
FAILURE MODE	Not applicable in the usual sense, because the arrangement does not act and therefore cannot malfunction. Its consequence rather than its failure is that every instructional control elsewhere in the catalogue becomes advisory, since nothing in the runtime interposes between an instruction and its execution.
FAIL STATE	fail-open
EMITTED SIGNAL	The flag is visible on each process command line and in the stored session record, so the posture is observable by an operator who looks for it; nothing announces it, and the warning that would normally announce it is explicitly suppressed in the settings file. The signal it would have to emit is a per-session posture record surfaced in the session listing and re-asserted at each dispatch, so that operating without prompts is a stated condition rather than a background fact.
METRICS	MG-031 MG-042 MG-050 MG-058
COST	low
FRICTION	low
AWS BINDING	Not applicable as a control. The equivalent cloud posture is running every workload under an unrestricted administrator role with no permission boundary and no Service Control Policy, which is precisely the arrangement this row catalogues.
AZURE BINDING	Not applicable as a control. The equivalent posture is an Owner role assignment at subscription scope with no deny-effect policy initiative above it.
ON HOST	yes
EVIDENCE	EV-020 EV-028 EV-032 EV-038
RESEARCH VALUE	The most important single row in the register for the thesis, because it is the mechanism by which every other instructional control on the host is reduced to advice. It is also the estate own honest statement of the trade it has made, unattended throughput bought by

deleting the only per-action human decision point, and as an experimental variable it is the baseline arm against which the prohibition entry is measured.

GR-L04-11 Concurrency cap on live agent sessions

INTENT	Bound the number of agents acting at once, so that the fleet aggregate rate of change stays inside one operator capacity to observe and to halt.
MECHANISM	Would be: a cap enforced at the creation chokepoint, refusing a new session when the count of live sessions in a scope, whether per product group, per host or per estate, is already at the limit, with an explicit queue and an operator override that is itself recorded as an event. The update plane demonstrates the shape one level up, with per-tier wave concurrency of one for production and one for the orchestrating host itself, but nothing anywhere bounds the number of concurrent agent sessions.
ENFORCEMENT	conditional
BINDS	AC-002 AC-012 AC-014 AC-019
ENABLES	II
PRECONDITION	A single creation chokepoint. There are presently seven distinct ways an agent comes into existence on this host and only two pass through the console, so a cap applied at the console binds two of seven paths.
BYPASS	Routed around by the six creation paths that never consult the console: the fleet wrapper started by the service manager, a raw multiplexer session from any shell, the console startup replay of stored desired sessions, the headless run endpoint, the browser terminal server, and the model-callable spawn tool in the agent runtime, which starts a fully autonomous headless session with permissions disabled and no human gate at all. A cap must therefore live at the process layer, as a task ceiling on a per-agent slice or as a lease held by a supervisor with an identity the agent lacks, rather than at the application interface. Even bound only to the console it still buys something real, namely that operator-initiated fan-out becomes a deliberate act with a visible count.
FAILURE MODE	A cap that counts sessions rather than acting agents will refuse creation while idle sessions occupy the count. Because the estate has no idle timeout at all, and one session has been idle for seven days, the cap would fill with dead sessions and block legitimate work, which is the failure most likely to get it removed.
FAIL STATE	fail-closed
EMITTED SIGNAL	A refusal event on creation naming the current and maximum counts, plus a continuous gauge of live sessions per scope. That gauge is directly the concurrent supervised stream count, which no channel on this host presently reports.
METRICS	MG-002 MG-044 MG-049 MG-052
COST	low
FRICTION	moderate
AWS BINDING	Lambda reserved concurrency, ECS service desired count with an upper bound, Step Functions Map state maximum concurrency, and Service Quotas as the account-level floor.
AZURE BINDING	Container Apps scale rules with a maximum replica count, Durable Functions fan-out concurrency limits, and Azure Batch pool quotas.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The direct manipulation for the attention half of supervision viability. Holding the brief constant and varying the cap from one to fourteen produces the dose-response curve relating concurrent supervised stream count to intervention latency, which is the empirical form of the claim that human oversight does not scale linearly with fleet size.

GR-L04-12 Advisory lock on scheduled automations

INTENT	Prevent a scheduled job from overlapping a still-running instance of itself, so that concurrent invocations cannot corrupt shared state or pile up unbounded.
MECHANISM	Every site-authored entry in the account schedule table is wrapped in a non-blocking advisory file lock, in most cases combined with a hard timeout, so a run still in progress when the next tick arrives is skipped rather than started. The per-minute automations carry a 50 second timeout against a 60 second period; the update plane per-minute tick carries an 890 second timeout under its own lock. The fleet registry writer uses the same primitive in blocking form on a separate zero-byte sentinel file, so that concurrent registry mutations serialise rather than being dropped, and it writes to a temporary file and renames. The scheduler additionally keeps per-agent per-day stamp files so that a matched schedule entry fires exactly once.
ENFORCEMENT	conditional
BINDS	AC-009 AC-012 AC-017 AC-019
ENABLES	II
PRECONDITION	Every writer must voluntarily take the same lock on the same path. There is no mandatory locking, and no file permission barrier that would make the agreement necessary rather than optional.
BYPASS	Advisory locking constrains cooperating processes and nothing else. Any of the agents, all of which run as the account that owns both the sentinel and the data, can invoke the same script without the wrapper or write the target document directly, and neither the lock nor the atomic rename convention will notice. In the registry case a direct truncate-and-write additionally destroys the atomic rename property, so the eight lock-free readers, which include the wrapper, the reload, kill, delete and unit creation scripts, the scheduler, the self test and the console own reader, can observe a torn document. Nothing detects a bypass: there is no checksum, no filesystem notification audit and no append-only journal, and the only forensic trace is a fifteen minute automatic commit into a repository the same account can rewrite. What it still buys is that the overlap failure mode, which is the one that actually occurs in practice, is eliminated for the honest path, and the non-blocking form degrades a slow run into a skipped tick rather than a pile-up.
FAILURE MODE	A lock file held or pre-created by an unrelated process silently suppresses the automation, and because the non-blocking form exits quietly a permanently held lock is indistinguishable from a job that had nothing to do. The estate has no contention counter and no alert on a lock held across many ticks.
FAIL STATE	fail-closed
EMITTED SIGNAL	Each wrapped job appends to its own log, so a skipped tick is inferable from a gap, and the registry writer exits non-zero with an explicit message when the lock cannot be taken. There is no contention counter and no event distinguishing a write that went through the sanctioned path from one that did not, which is the event the control actually needs.
METRICS	MG-014 MG-060 MG-069
COST	low
FRICTION	low
AWS BINDING	EventBridge Scheduler with a single-invocation target plus a DynamoDB conditional write as a distributed lease, Step Functions Standard workflows whose execution names give server-side uniqueness, and Systems Manager Change Calendar for the schedule itself.
AZURE BINDING	A Blob lease as a distributed mutex, Durable Functions singleton orchestration on a fixed instance identifier, and Logic Apps concurrency control set to one.
ON HOST	yes
EVIDENCE	EV-022 EV-027 EV-030 EV-031
RESEARCH VALUE	A rare case where the estate convention is correct in design and unenforceable in mechanism, which makes it the clearest illustration in the catalogue of the difference between a discipline and a control. Replacing the advisory lock with a lease held by a separate identity converts the identical guarantee from conditional to structural without

> changing any application logic, so the pair prices enforcement independently of the rule being enforced.

2.5 L5. Plano de dados

Esta camada governa o que o estado persistido de um agente, e qualquer sistema de registro upstream ao qual ele se conecte, podem representar ou permitir de modo structural. Ela contém o exemplo mais claro do catálogo de um controle que hoje é genuinamente structural: a identidade de conexão com o System of Record é somente leitura, de modo que uma escrita não é uma permissão que falta ao agente, mas um estado que essa identidade simplesmente não consegue expressar `REPORTED`. Essa é a única entrada nesta camada, e uma de apenas duas em todo o catálogo, que é ao mesmo tempo presente e structural.

Tudo o mais nesta camada é conditional ou está ausente. O registro da frota usa um único escritor com um lock consultivo (advisory lock), presente mas conditional, porque um advisory lock é uma convenção que cada escritor precisa optar por respeitar, e a conta que executa cada escritor é a mesma conta `EV-022`. Proveniência em nível de campo, restrições de esquema que tornam um estado proibido irrepresentável, e a negação ao agente da credencial que poderia alterar seu próprio registro de auditoria são todas propostas e todas structural; juntas, são o que estenderia o padrão de conexão somente leitura com o System of Record de uma única conexão para todo repositório que este catálogo nomeia, incluindo o repositório de evidência discutido na L8.

Camada 5. Plano de dados.

GR-L05-01 Read-only connection identity to the System of Record

INTENT	Ensure the upstream System of Record cannot be modified by any agent-driven process, by binding the connection itself to an identity that has no write capability.
MECHANISM	Realised in two forms, only one of which is verifiable from this host. On the upstream side the connector is reported to point at a read-only reporting replica rather than the live tenant, to be retrieval-only with no write helper present in the module at all, to refuse the vendor read-write contract interface path, to clamp result sizes and to refuse paging of large collections; that connector runs on the Platform host and is reported here, not observed. What is observed on this host is the identical pattern realised twice. The update plane splits its transport into a read function that accepts any declared host descriptor and an apply function that accepts only a branded target type, so read and write are different call signatures rather than the same call with a flag. The agent runtime exposes seven read-only tools, covering cloud instance description, repository read and search, knowledge search and read, and runtime information, none of which has a write counterpart anywhere in the module, and the local keyword search service over the knowledge corpus is likewise search-only.
ENFORCEMENT	structural
BINDS	AC-038 AC-039 AC-051 AC-052
ENABLES	II
PRECONDITION	The upstream system must expose a replica or a role that is read-only at the server, rather than a client that merely declines to write.
BYPASS	REPORTED from the founding case, not verified on this host: no write-capable identity is issued for the System of Record at all, so a write is not expressible through this connection regardless of what a client requests. The mechanism sits outside the Agent Console host, so the specific server-side implementation, a role grant restricted to select, a physically separate read replica, or another architecture, was not verified in this run; what is reported is the absence of a write-capable credential, not the particular implementation that enforces it. See F-023.
FAILURE MODE	A connector repointed from the replica to the live tenant by a configuration change is indistinguishable from a correct one until the first write is attempted, and the credentials for both are reported to live in the same configuration table; the failure is therefore a silent change of blast radius rather than an error.
FAIL STATE	fail-closed
EMITTED SIGNAL	None today on this host. Reads produce ordinary application logs and there is no event distinguishing a read-only connection from a read-write one, nor any assertion at connection time. The signal it would have to emit is a connection establishment record naming the resolved identity and its capability, plus a server-side refusal event whenever a write is attempted against the replica.
METRICS	MG-030 MG-031 MG-033 MG-043
COST	low
FRICTION	moderate
AWS BINDING	RDS read replicas with a database role granted only SELECT, IAM database authentication so the identity is minted per session rather than stored, and Lake Formation or Redshift read-only grants for the analytic surface.
AZURE BINDING	Azure SQL read-only replicas reached with a read-only application intent, a database principal in the data reader role only, and Microsoft Entra authentication so the connection identity is not a shared secret on disk.
ON HOST	partial
EVIDENCE	REPORTED
RESEARCH VALUE	The only data-plane control in the estate that could be made structural with no change to agent behaviour at all, because the enforcement moves entirely to the server. It is therefore the ideal test of whether structural enforcement changes measured agent behaviour: the prediction is that boundary crossing success rate falls to zero for write classes while task completion rate is unchanged, because the agent never needed the write in the first place.

GR-L05-02 Single writer with advisory lock on the fleet registry

INTENT	Guarantee that fleet state is mutated through one validated, serialised and atomic path, so that a console edit and a scheduled reload cannot corrupt or clobber each other.
MECHANISM	One shell script declares itself the only writer of the registry document and every mutation funnels through a single write function. That function opens a descriptor on a separate zero-byte sentinel file, takes a blocking advisory lock, applies a jq program with bound string arguments, writes to a process-suffixed temporary file in the same directory and renames it over the target, then releases. Because values arrive as bound jq arguments, no value can inject document structure. Validation gates live inside the writer: the agent must not already exist, the product group must exist, the tool must be marked wired in the catalogue, the model must be listed, status is a closed enum of active or retired, schedule times must match a strict pattern, and a group that still has agents cannot be removed. The console never writes the document; it shells out to the script with an argument array and adds its own name and identifier validation plus request schemas at the route boundary.
ENFORCEMENT	conditional
BINDS	AC-017 AC-018 AC-020 AC-021
ENABLES	II
PRECONDITION	Every writer must agree to use the script. Nothing in the filesystem or the privilege model requires that agreement.
BYPASS	The single-writer property is a convention with no enforcement beneath it. The document is mode 644 owned by the account every agent runs as, so any agent opens it for writing directly and never touches the sentinel, and the lock is advisory so it binds only processes that ask for it. A truncate-and-write additionally destroys the atomic rename property, so the eight lock-free readers can observe a torn document. The account also holds passwordless root, so even a root-owned mode 600 registry would not constrain it. Nothing detects a bypass: no checksum, no filesystem notification audit, no append-only journal, and the only trace is a fifteen minute automatic commit into a repository the same account can rewrite. There is also a genuine defect inside the honest path: the reload subcommand re-reads the registry after releasing the lock in order to emit the launch descriptor, so a concurrent staging of a different model can make that descriptor disagree with the registry, and the wrapper prefers the descriptor, which means the race decides which model an autonomous agent boots on. What it still buys is that every caller who does use the script inherits the validation gates and the atomic rename automatically, which is why the console can be a thin caller and still be safe.
FAILURE MODE	A failing jq program leaves the original untouched and removes the temporary file, which is correct. The dangerous failure is silent schema drift: the shape is enforced only by the jq programs inside one script, there is no schema document, no migration mechanism and no runtime validator, and the console typed interface is a cast over a parse rather than a check, so a malformed entry flows into the console typed and wrong.
FAIL STATE	fail-closed
EMITTED SIGNAL	The writer exits non-zero with an explicit message on lock failure or validation failure, and the fifteen minute automatic commit records the resulting difference. There is no event that distinguishes a write made through the script from one made around it, which is exactly the event the control requires in order to be evaluable.
METRICS	MG-059 MG-063 MG-064 MG-065
COST	low
FRICTION	low
AWS BINDING	DynamoDB with conditional writes and optimistic concurrency on a version attribute so serialisation is server-side, or RDS with a transaction and a unique constraint, with S3 Object Lock where the record must additionally be immutable.
AZURE BINDING	Azure Table Storage or Cosmos DB with entity tag optimistic concurrency, a Blob lease as the mutex, or Azure SQL with a transaction, plus a Blob immutability policy for the retained copy.
ON HOST	yes
EVIDENCE	EV-017 EV-020 EV-022

RESEARCH VALUE	The best-argued convention in the estate and therefore the sharpest test of the thesis: the design is correct, the code is careful, and the guarantee is still unavailable because the substrate does not require it. Moving the same schema to a store with server-side conditional writes changes nothing about the application and converts the fail closed ratio for this class from a claim into a property.

GR-L05-03 Connection identity that structurally cannot write

INTENT	Make prohibited writes impossible at the server rather than declined at the client, so that the calling process compliance is not a term in the outcome.
MECHANISM	Would be: provision a distinct database or interface principal per agent role whose server-side grants contain no insert, update, delete or schema definition privilege on any object, minted just in time with a short session lifetime rather than issued as a long-lived secret, and reachable only over a private endpoint so the credential is useless from anywhere else. Application code then becomes irrelevant to the guarantee, because a write returns a permission error from the server whatever the client intended.
ENFORCEMENT	structural
BINDS	AC-027 AC-039 AC-051 AC-052
ENABLES	I
PRECONDITION	The upstream system must support per-principal grants, and the agent must hold no other credential to the same system. Today five estate database credentials and superuser material sit on disk under the single account, so a scoped principal would coexist with unscoped ones and inherit none of their constraint.
BYPASS	The control is not defeated by beating the grant; it is defeated by finding another credential. On this host that is trivial, because the secrets directory holds credentials for five estate databases and the update plane preflight already reaches a database through a local superuser role. It is further defeated by lateral reach, since private keys for seven fleet hosts sit on the same disk and a write can simply be performed from the far side of a remote shell. The control is therefore only as strong as the credential inventory surrounding it, which makes credential removal its precondition rather than its complement; scoping one identity while leaving an unscoped one adjacent produces a boundary that exists on paper and not in the process.
FAILURE MODE	A grant that omits one object, a sequence, a view base table or a schema added later, produces intermittent write success. That is worse than no control, because it creates a justified belief in a boundary that holds most of the time.
FAIL STATE	fail-closed
EMITTED SIGNAL	A server-side permission denied event per attempted write, naming the principal and the object, on the database own audit stream. This is an infrastructure-emitted signal independent of the agent, and it is directly a boundary crossing attempt rather than a proxy for one.
METRICS	MG-031 MG-032 MG-033 MG-040
COST	moderate
FRICTION	moderate
AWS BINDING	IAM database authentication against an RDS read replica with a role granted only SELECT, an explicit deny on connect for every other principal, S3 bucket policies and Object Lock for object stores, and Lake Formation permissions for analytic reads.
AZURE BINDING	Microsoft Entra authentication to Azure SQL with the principal in the data reader role only, read-only replica routing, Blob immutability, and container-scoped shared access signatures limited to read.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The purest available instance of the thesis central distinction, because it removes agent compliance from the outcome entirely while changing nothing the agent can perceive except the error it receives. It is the benchmark against which every prose-level prohibition in the model and prompt layer should be scored, since the two express the same rule with and without enforcement.

GR-L05-04 Schema constraints making a prohibited state unrepresentable

INTENT	Express governance rules as data constraints so that a forbidden combination of values cannot be stored at all, rather than being prevented by whichever code path happens to check.
MECHANISM	Would be: encode as declarative constraints on the store the rules that presently live in procedural checks. A check constraint requiring that a record marked approved carries a non-null approver distinct from the submitter; a foreign key with restricted delete on every audit event; a unique partial index preventing two concurrent runs against one target; an enumerated type rather than a free string for status; a not-null deployed revision on any record in a deployed state. The estate already contains the idea in one place and in a different form: a transition table in which the report-only lane has an empty outgoing edge set for every status, so no route handler can promote such a record whatever its code does.
ENFORCEMENT	structural
BINDS	AC-017 AC-054 AC-055 AC-056 AC-058
ENABLES	II
PRECONDITION	A store that supports declarative constraints. This host has no relational database at all; every piece of state is flat file JSON, and the only shape enforcement anywhere is a set of jq programs inside one shell script.
BYPASS	A constraint binds writes made through the store and nothing else. It is routed around by writing out of band, which on this host is trivial because the state is a file owned by the agent account, and which remains possible on a relational store for any principal holding schema definition rights, since dropping a constraint is itself a write the same credential can perform. The durable form therefore separates the schema-changing identity from the data-changing identity and treats a constraint change as a reviewed migration. A subtler bypass is representational rather than technical: a rule expressed as a constraint on one table is satisfied by writing a semantically equivalent record elsewhere, which is exactly what happens when a peer review is recorded under the reviewer credential by the party that authored the change.
FAILURE MODE	A constraint added without backfilling existing rows fails its migration and is quietly dropped from the change set, leaving the belief that it exists; a constraint that encodes a rule the organisation later relaxes becomes an outage rather than a warning, which is the pressure that gets constraints removed.
FAIL STATE	fail-closed
EMITTED SIGNAL	A constraint violation error at write time carrying the constraint name, on the store own error stream, plus a periodic drift report comparing declared and live definitions. The estate demonstrates why the second is necessary, since its declared schema has already drifted behind the live database it purports to describe.
METRICS	MG-031 MG-036 MG-060 MG-067
COST	moderate
FRICTION	moderate
AWS BINDING	RDS check constraints, foreign keys with restricted delete and enumerated domains; DynamoDB condition expressions for the same effect in a key-value store; CloudFormation change sets with drift detection to keep declared and live definitions aligned.
AZURE BINDING	Azure SQL check constraints and foreign keys, Cosmos DB conditional writes through entity tags and stored procedures, and Azure Resource Manager deployment stacks with deny assignments and what-if analysis to detect drift.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Tests whether governance expressed as data outperforms governance expressed as procedure when the procedure is written by the governed party. The estate supplies its own comparison: the empty transition table rule has never been violated, while the reviewer independence rule, which exists only as prose and has no code check anywhere, has already been violated in production.

GR-L05-05 Separation of the evidence store from the operational store

INTENT	Hold the record of what an agent did somewhere the agent ordinary operation cannot reach, so that the record survives the agent and the incident.
MECHANISM	Would be: ship every audit and transcript event at write time to a store in a different trust domain, with a different credential, append-only semantics and a retention lock the writing principal cannot shorten. The agent principal holds put permission only, with no list, get, overwrite or delete, so it can add to the record and can neither read nor revise it. Sequence numbers at the source allow gaps to be detected at the destination.
ENFORCEMENT	structural
BINDS	AC-022 AC-053 AC-054 AC-055 AC-058
ENABLES	II
PRECONDITION	A second trust domain, which the founding case does not have on this host: the console source, its compiled output, its environment file and its audit log all sit in a tree owned by the same account as the agents, and about 1.13 GiB of model runtime transcripts sit on the same ordinary filesystem.
BYPASS	Separation is defeated at the point of production rather than at the point of storage. If the event is generated by a process the agent controls, the agent can decline to generate it, and no downstream immutability repairs a record that was never written; the estate demonstrates this precisely, since its console audit log records what operators did through the console and records nothing whatever that an agent does. It is defeated a second way by credential reach: the shipping credential must not be readable by the agent, which fails immediately here because every secret is a file under the one account. The control therefore becomes real only once the emitting process is outside the agent control, which makes it dependent on an infrastructure-emitted event stream rather than a substitute for one.
FAILURE MODE	A shipping pipeline that buffers locally and fails quietly produces a store that looks complete and has holes exactly where the incident is, because the conditions that cause incidents, disk pressure, a restart, a killed process, are the same conditions that lose the buffer.
FAIL STATE	fail-open
EMITTED SIGNAL	A delivery receipt per event and a continuous gap detector comparing sequence numbers at source and destination. The absence of an expected event is itself the signal, which is why sequencing is mandatory rather than an optimisation.
METRICS	MG-048 MG-064 MG-065 MG-066
COST	moderate
FRICTION	low
AWS BINDING	CloudTrail with an organisation trail delivering into a separate log archive account, S3 Object Lock in compliance mode with a bucket policy denying delete to every principal, and CloudWatch Logs cross-account subscription filters for the streaming half.
AZURE BINDING	Azure Monitor diagnostic settings shipping to a Log Analytics workspace in a separate subscription, immutable Blob storage with a time-based retention policy, and a separate tenant boundary where the separation must survive subscription compromise.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Attacks the founding case headline measurement directly, the 61.1 per cent factual divergence between agent self-report and independent evidence across 18 audited episodes, by giving the independent evidence a home the agent cannot touch. Computing claim divergence rate with and without the separated store measures how much divergence was previously unmeasurable rather than absent.

GR-L05-06 Agent denied the credential that could alter its own audit record

INTENT	Remove from the agent reach any credential whose use would let it rewrite, truncate or suppress the record of its own actions.
MECHANISM	Would be: a credential inventory keyed by what each secret can do to the evidence plane, plus a rule that the agent runtime environment contains none of them. The audit sink accepts only an append-scoped, write-only token; the log shipping identity is held by a sidecar or host agent running under a different account; the administrative credential for the store exists only behind a break-glass path with its own approval. The agent own token is minted per session and expires with it, so a leaked copy is worthless outside its window.
ENFORCEMENT	structural
BINDS	AC-027 AC-053 AC-054 AC-055 AC-058
ENABLES	II
PRECONDITION	Per-agent identity and a place to keep credentials the agent account cannot read. Neither exists: every agent runs as one account, that account secrets directory holds tokens for the notification bridge, the update plane, the model provider and six deploy gate identities, and a file mode of 600 provides no separation when every process shares the same user identifier.
BYPASS	On this host the control is unbuildable as stated, because the account holds unrestricted passwordless root, so every file mode, every ownership and every other account secret is reachable by definition. Even after the root grant is removed the practical bypass is credential adjacency rather than credential theft: the estate already normalises one agent using another agent credential, since a scheduled watcher runs on behalf of one agent while presenting a second agent deploy gate token, and a recorded incident has a session reading a peer working directory and using the peer token to record a formal review of its own work. Denying the audit credential specifically is therefore necessary and not sufficient; the fleet whole credential-sharing habit has to change alongside it, which makes this control a programme rather than a setting.
FAILURE MODE	A partial removal leaves one path still carrying the credential, whether an environment variable inherited by a spawned child, a value compiled into source, or a copy inside a backup directory, and the belief that it has been removed is more dangerous than knowing it is present. The estate contains exactly this shape already: a plaintext authentication token compiled into a world-readable source file guarding a wildcard-bound socket that spawns a fully privileged session.
FAIL STATE	fail-closed
EMITTED SIGNAL	Every use of an evidence-plane credential logged at the sink with the presenting principal, so that a use by an unexpected principal is visible, plus a periodic scan asserting that no evidence-plane credential is present on any agent-reachable path.
METRICS	MG-029 MG-042 MG-065 MG-067
COST	moderate
FRICTION	low
AWS BINDING	IAM policies granting only log put or object put with an explicit deny on delete and get, credentials delivered through an instance or task role rather than a file on disk, and Secrets Manager resource policies scoping who may retrieve what.
AZURE BINDING	Managed identity holding only a publisher or append role, Key Vault access policies scoped per identity, and immutable Blob containers so that delete is not a grantable permission at all.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Isolates the variable that makes audit records evidential rather than merely descriptive. Because the estate has a dated separation-of-duties bypass performed with a peer real credential rather than a hypothetical one, this control can be evaluated against an actual counterfactual, which is rare in governance research.

GR-L05-07 Atomic write discipline for flat file state

INTENT	Ensure a reader of shared state never observes a partially written document, without requiring coordination between reader and writer.
MECHANISM	Write to a temporary file in the same directory then rename over the target, so replacement is a single rename on one filesystem. Applied in at least three places: the fleet registry writer uses a process-suffixed temporary file and renames inside its lock; the update plane persistence module states the rule in its own header and applies it to its scan, proposal and run documents; and the launch descriptor handed to the agent wrapper is written the same way. The update plane audit stream is a separate append-only line-delimited file rather than a rewritten document, which is the right shape for a record that must only grow.
ENFORCEMENT	conditional
BINDS	AC-017 AC-023 AC-053 AC-058
ENABLES	II
PRECONDITION	Temporary file and target must share a filesystem, and every writer must follow the pattern; nothing requires either, and no test asserts either.
BYPASS	The property protects honest readers from honest writers and protects nothing from a writer that does not participate. A direct truncate-and-write from any agent, which the account can perform on every one of these documents, defeats it at once, and the readers are the exposed party precisely because none of them takes the lock. It is silently defeated a second way by a temporary file placed on a different filesystem, at which point the rename degrades to a copy and atomicity is lost with no error anywhere. And atomicity is not durability: none of these writers synchronises the containing directory, so a power loss can lose a rename that appeared to succeed. What it still buys is genuine: it removes the torn-read failure mode across every scheduled reader on the box, which matters because eight distinct readers parse the fleet registry with no coordination at all.
FAILURE MODE	The launch descriptor is written atomically but is covered by no lock, so two concurrent reloads for one agent race with only the rename for protection, and the winner determines which model an autonomous agent boots on. That is a correctness guarantee sitting one level below the decision it is protecting.
FAIL STATE	fail-closed
EMITTED SIGNAL	None today. A successful rename emits nothing and a downstream parse failure is the first observable symptom. The signal it would have to emit is a write generation counter or content hash per document, recorded at write time and checked by readers, which would make an out-of-band modification detectable rather than invisible.
METRICS	MG-060 MG-064 MG-065
COST	low
FRICTION	low
AWS BINDING	S3 put object with strong read-after-write consistency plus versioning, and Object Lock where the prior version must survive; DynamoDB conditional writes where the state is record-level rather than document-level.
AZURE BINDING	Blob storage put blob with conditional headers and blob versioning, Cosmos DB entity tag conditional replace, and a Blob immutability policy where prior versions must be preserved.
ON HOST	partial
EVIDENCE	EV-022 EV-031
RESEARCH VALUE	Demonstrates that correctness engineering and governance enforcement are separable properties of the same mechanism: the rename gives a real correctness guarantee and no governance guarantee whatever. Adding a content hash converts it into a tamper-evidence primitive at almost no cost, which makes it the cheapest available intervention on audit record immutability in the whole catalogue.

GR-L05-08 Relational store with transactional integrity for fleet state

INTENT	Give fleet state the properties a governance record requires: atomic multi-record change, server-enforced constraints, isolation between concurrent writers, and a query surface that does not depend on parsing a file.
MECHANISM	Would be: move the registry, the desired session set, session metadata, resource accounting and the update plane proposals and runs into a relational database with explicit transactions, so that a state change and its audit event commit together or not at all. The deploy gate on the Platform host already demonstrates the pattern the fleet lacks: its status changes run as atomic conditional transitions with the prior status in the predicate, so that a check-then-act race returns a conflict rather than a double transition, and its event table carries a restricted-delete foreign key so events cannot be orphaned or removed.
ENFORCEMENT	structural
BINDS	AC-017 AC-018 AC-055 AC-056 AC-058
ENABLES	II
PRECONDITION	A database and somewhere to run it. There is no relational store on this host at all; every piece of state is flat file JSON, and the update plane persistence module states the absence as a design premise rather than an omission.
BYPASS	A transactional store enforces integrity against writers that connect through it and enforces nothing against a writer with filesystem access to its data directory or with schema definition rights on its objects. On this host, where the account holds passwordless root, an agent could stop the database, edit its files and restart it, so the store has to live on a host the agent does not administer for the guarantee to hold at all. There is a second and subtler bypass the estate has already experienced at the application layer: a transaction guarantees that the recorded state changes atomically and guarantees nothing about whether the recorded state describes reality, which is exactly how an approved flag can coexist with a running process serving unapproved code.
FAILURE MODE	A transaction boundary drawn around the state change but not around the audit write reintroduces the gap it was meant to close. Separately, an unavailable database converts a previously local operation into a distributed failure, so every fleet supervision script would need a defined behaviour when the store is unreachable, and today none has one.
FAIL STATE	fail-closed
EMITTED SIGNAL	Commit and rollback events, constraint violations and conflict responses on concurrent transitions, all emitted server-side and therefore independent of the process that attempted the change.
METRICS	MG-060 MG-063 MG-064 MG-067
COST	high
FRICTION	moderate
AWS BINDING	RDS or Aurora PostgreSQL with per-role grants and IAM database authentication, DynamoDB transactions where a key-value shape suffices, and automated backups with point-in-time recovery for the state history.
AZURE BINDING	Azure Database for PostgreSQL flexible server or Azure SQL with Entra authentication and per-role grants, Cosmos DB transactional batch operations, and continuous backup.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Allows the identical governance rules to be run twice, once over flat files with advisory locks and once over a transactional store, with everything else held constant. That is the sharpest available measurement of how much audit record completeness and fail closed ratio are attributable to the substrate rather than to the rules written on top of it.

GR-L05-09 Backup and point-in-time recovery for agent state

INTENT	Make agent-managed state recoverable to a known-good moment, so that a destructive or mistaken change becomes a restore rather than a reconstruction.
MECHANISM	Would be: scheduled, versioned, off-host backups of the fleet registry, session metadata, the instruction corpus and the evidence stores, with a stated recovery point objective, an immutable retention window and a restore procedure rehearsed on a schedule rather than assumed. The update plane already implements the equivalent for host state, taking a cloud snapshot before any change and refusing to proceed unless every attached volume reports complete, retaining artefacts seven days and pruning only what it created; the gap is that no equivalent exists for the agent plane own state.
ENFORCEMENT	conditional
BINDS	AC-002 AC-017 AC-023 AC-024 AC-054
ENABLES	II
PRECONDITION	A defined state inventory. Today state is scattered across flat files in at least four trees, plus a git repository that an automatic commit rewrites every fifteen minutes, plus transcripts under the model runtime own directory, and no single document enumerates them.
BYPASS	A backup does not prevent an action, it only bounds its cost, so it changes the reversibility of an action class rather than its reachability, and the register should not credit it with containment. It is defeated by scope, when the backup omits the one object that mattered; by adjacency, when the copy is written under the same account a destructive agent controls; and by the recovery gap between the last good copy and the incident. On this host the third is uneven: the transcript backup runs every five minutes with no delete flag, which is generous, while the fleet registry only history is a repository the same account can rewrite. The estate also shows the honest limit of snapshot-based recovery, since its update plane has whole-machine snapshots and no package state capture at all, so a single-package rollback is reconstructed from a command line rather than restored from a copy.
FAILURE MODE	An untested restore. A regime that has never been exercised fails at the only moment it matters, and the failure is discovered then; the estate snapshot path has never fired in production, so its restore behaviour is entirely unproven.
FAIL STATE	fail-open
EMITTED SIGNAL	A backup completion receipt per protected object with a verified recovery point, plus a periodic restore rehearsal result. The receipt must record verification rather than submission, which is the property the update plane snapshot receipt already gets right by minting only after the provider reports completion.
METRICS	MG-018 MG-021 MG-038 MG-043
COST	moderate
FRICTION	low
AWS BINDING	AWS Backup with vault lock, scheduled EBS and instance snapshots, S3 versioning with lifecycle rules and Object Lock, and RDS automated backups with point-in-time restore.
AZURE BINDING	Azure Backup with immutable vaults and soft delete, policy-driven managed disk snapshots, Blob versioning with point-in-time restore, and Azure SQL long-term retention.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Converts reversibility from a property asserted in a taxonomy into a measured one. Rehearsing restores gives the rollback exercise success rate real values, which is the only honest way to test whether the reversibility ratings assigned to the sixty action classes in this register are accurate rather than optimistic.

GR-L05-10 Field-level provenance on every state record

INTENT	Make every stored value carry who set it, when, through which mechanism and under what authority, so a record can be audited without reconstructing the process that produced it.
MECHANISM	Would be: extend each state record with per-field provenance, namely actor identity, actor type distinguishing human from agent from automation, timestamp, the originating request or run identifier, and the configuration hash in force at the time; and write provenance in the same transaction as the value so the two cannot diverge. The deploy gate on the Platform host shows the partial version, recording who decided and when on the record itself plus an append-only event row per action, with the actor set by the middleware from the authenticated session and never taken from the request body.
ENFORCEMENT	structural
BINDS	AC-017 AC-053 AC-055 AC-056 AC-058
ENABLES	II
PRECONDITION	An identity to attribute writes to. With one shared operating system account and no per-agent identity, provenance would record the same actor for every write, which is the reason the fleet registry carries no actor field at all today.
BYPASS	Provenance is only as strong as the identity it records and the place it is written, and the estate contains a live example of each weakness. Where the actor is derived from a self-asserted field it can be set to anything: an operator permission array is taken from a request body and used for every permission check, and the notification bridge source name is free text the caller chooses, so both are impersonable by construction. Where the actor is derived from a presented credential it is defeated by using a peer credential, which has already occurred on the deploy gate. And where value and provenance live in the same agent-writable file, both are rewritten together in one edit. The structural version therefore requires the server to set the actor from an authenticated principal, the record to be append-only, and the store to be outside the agent reach; provenance without those three is decoration.
FAILURE MODE	Provenance written asynchronously or on a best-effort basis produces records whose fields disagree with their own history, and the disagreement is invisible until an audit. A partially adopted scheme is worse than none, because it implies a coverage it does not have.
FAIL STATE	fail-open
EMITTED SIGNAL	The provenance fields themselves, queryable per record; the derived signal that matters for governance is the proportion of records whose actor is an agent rather than a human, which is directly the self-attestation share of the record and is not computable at all today.
METRICS	MG-029 MG-064 MG-066 MG-067
COST	moderate
FRICTION	low
AWS BINDING	DynamoDB item attributes written in the same transaction with the caller identity taken from the request context, CloudTrail for the control plane half, and Config for configuration state with a per-resource change history.
AZURE BINDING	Cosmos DB or Azure SQL columns populated from the authenticated principal, Activity Log for the control plane half, and Resource Graph change history with Change Analysis.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Provenance is the instrument that makes several other measurements possible rather than an intervention in its own right. Without it the self-attestation share of the record cannot be computed, episode reconstructability stays qualitative, and separation of duties violations can only be found by reading incident narratives, which is how the estate found its own.

2.6 L6. Mudança e implantação

Esta camada governa como uma mudança chega a um sistema em execução e quem ou o que a controla. Boa parte dela reafirma práticas que a entrega contínua estabeleceu bem antes de sistemas agênticos existirem: ordenação canário e por ondas (wave), dry run como padrão, snapshots antes da aplicação. Quatro entradas desse tipo são parciais neste host, presentes em código mas não em vigor de forma universal EV-030 EV-031. A única entrada da camada genuinamente structural e genuinamente presente é o modo de teste do domínio de tempo real, que intercepta todo despacho externo antes que um destino real seja resolvido, na camada de despacho e não no raciocínio do agente, de modo que a intenção do agente de alcançar um destino real é expressável e simplesmente não resulta em um REPORTED. É a segunda das duas entradas presente-e-structural do catálogo, ao lado de GR-L05-01, e ambas compartilham a mesma forma: o efeito proibido não é alcançável, independentemente do que o agente decida.

A própria aprovação de implantação é apenas parcial aqui, por um motivo específico e documentado. A autoridade de aprovação é structural, porque somente um token de identidade humana pode escrever um status aprovado, e nenhum token de agente alcança esse middleware EV-030. O que não é enforced é a vinculação entre uma revisão aprovada e o que de fato é executado, porque o gerenciador de processos reexecuta o que quer que esteja na árvore de trabalho compartilhada, independentemente do que foi aprovado; a mitigação para essa lacuna é uma instrução de prompt, motivo pelo qual GR-L06-11 proveniência de artefato de implantação permanece proposta em vez de presente.

Camada 6. Mudança e implantação.

GR-L06-01 Deployment approval gate held by a separate principal

INTENT
Ensure no new application code reaches a production service until a human principal, distinct from the agent that prepared it, records an approval.

MECHANISM
A gate on the Platform host, referred to internally as patchgate, whose complete source is readable from this host. A fleet agent submits a patch instead of deploying it; a lead agent may record a technical peer review, which the code returns as explicitly not an approval; one named human identity approves; only then may the deploy agent execute. Approval is a status field write plus an append-only event row inside a database transaction, performed as an atomic conditional transition with the prior status in the predicate, so a check-then-act race yields a conflict rather than a double transition. The approving identity is hardcoded rather than held in configuration, is required to be a local login rather than a federated one, and the actor is set by middleware and never read from the request body. The rollback target is resolved server-side from version control, pinned first-write-wins and never caller-supplied; automatic rollback takes no target parameter and proves the target is an ancestor of the current head. The risk classifier adjudicates the diff rather than the submitter declaration, fails closed on unparseable paths, and refuses to fast-lane any change to the gate itself. Six per-agent bearer tokens for this gate are stored on this host, and a per-minute watcher here polls the gate and types a prompt into the deploy agent session whenever the approved set changes.

ENFORCEMENT
conditional

BINDS
AC-004 AC-009 AC-043 AC-044 AC-045 AC-056

ENABLES
II

PRECONDITION
Separation of duties requires four things: distinct identities, distinct credentials, a mechanism by which one identity credential cannot be presented on another behalf, and an approval authority the submitting party does not hold. The first and fourth are present here; the second and third are not.

BYPASS
Approval itself is genuinely hard to bypass, and that should be credited: no agent bearer token can reach the approve route, which demands a human session for one hardcoded local identity with a live re-check that the account is not federated. Everything downstream is where it fails. First, separation of duties is defeated at the credential layer rather than the authority layer: all six per-agent gate tokens are mode 600 files owned by the one account every agent runs as, and every one was confirmed readable from this collection session, so any agent can present any agent identity. The estate own scheduled watcher already does this routinely, running on behalf of one agent while presenting another token, and a recorded incident has a session using the lead agent token to record a formal peer review of a patch it had itself originated; no code anywhere compares the originating agent against the reviewing agent, and the tokens have not been rotated since provisioning despite a documented recommendation. Second, and more fundamentally, the gate governs a record and not an artefact: approval flips a database field and nothing binds the running process to that field. The gated service executes from a shared working tree with no build step and no pinned artefact, so any restart, including the process manager own memory-triggered restart, activates whatever is on disk at that instant; this has happened at least twice, once shipping code hours before its deploy event was recorded. Third, on this host the production working tree is mounted read-write over a network filesystem, so an ordinary file write from this box places code the next restart will serve, needing no shell on the target, no gate call and no compromise of any trusted agent. The mitigation for the shared tree is a prompt instruction telling the model to use an isolated worktree, and the only technical response is a drift detector that deliberately performs no remediation. What the gate still buys is that approval authority is unreachable from any agent credential and that an append-only trail of intent exists.

FAILURE MODE
The gate stays healthy and correct while the running code diverges from the approved record; that is the failure its drift detector was written to notice after the fact, and the detector own message is that the running code may not be what was approved.

FAIL STATE
fail-open

EMITTED SIGNAL
An append-only event row per action covering submission, lead review, approval, cancellation, deploy start, deploy success, deploy failure and automatic rollback, plus a message to the operator on submission and a periodic drift comparison of the live

	revision against the last recorded deployed revision. The gap in the signal is decisive: no event is emitted when code reaches production by any path other than the gate, and that is the path which has actually been used.
METRICS	MG-007 MG-024 MG-051 MG-067
COST	moderate
FRICTION	high
AWS BINDING	CodePipeline manual approval actions with the approval permission granted to a principal distinct from the pipeline role, CodeDeploy consuming an immutable revision from S3 so the deployed bits are bound to the approval, and Systems Manager Change Manager for the change record itself.
AZURE BINDING	Azure DevOps or GitHub environment approvals with required reviewers drawn from a different group than the pipeline identity, deployment gates on health signals, and Azure Resource Manager deployment stacks so the deployed definition is the reviewed definition.
ON HOST	partial
EVIDENCE	EV-027 EV-028 EV-030 EV-032
RESEARCH VALUE	The estate most instructive control, because it is simultaneously the best engineered and among the least effective. It separates two properties usually conflated, authority over a decision record and control over the artefact the decision was about. Measuring separation of duties violations and verification independence against it, before and after binding deploys to an immutable artefact, tests whether approval regimes derive their value from the approval or from the binding.

GR-L06-02 Pre-apply snapshot of the target state

INTENT	Guarantee that a recovery point exists, and is confirmed complete, before any change is applied to a target.
MECHANISM	In the update plane a snapshot module is the only place that can mint a snapshot receipt, and it mints one only after the cloud provider reports the snapshot available; for volume-backed instances every attached volume must reach the completed state, on the stated principle that a partial set is not a recovery point. The apply function signature demands a receipt, so a host whose snapshot failed is not merely skipped by a check, it is uncallable, and the failure path records that nothing was applied because the snapshot failed. Polling runs at fifteen second intervals with a thirty minute ceiling, artefacts are retained seven days, and pruning removes only artefacts the module itself created. The scope gap is explicit: there is no package state capture, no package manager history capture and no dependency state dump anywhere in the module, so below the whole-machine snapshot the only rollback affordance is a reconstructed downgrade command.
ENFORCEMENT	conditional
BINDS	AC-024 AC-031 AC-043 AC-044 AC-045
ENABLES	II
PRECONDITION	A provider snapshot interface and a declared inventory naming the instance or volume identity to snapshot; the plane inventory is hand-maintained and explicitly not discovered at runtime by network scan, host configuration or any live cloud listing.
BYPASS	Within a given compiled build the requirement is not evadable by forgetting, because the apply function cannot be called without a receipt object that only the snapshot module can create; that is a stronger property than a runtime check, since no code path exists that omits it. It is defeated one level out. The compiled output on this host is mode 644 owned by the agent account, so an agent rewrites the module and restarts the console process; the plane state directory and its inventory file are likewise agent-writable, so tier, enabled flag, credential path and restart constraints can all be edited. And the control binds only changes made through this plane: a change made by a direct shell session, by the model-callable remote execution tool, or by an ordinary file write to the read-write mounted production tree takes no snapshot at all and is invisible to this mechanism. Finally a whole-machine snapshot is a coarse instrument, so in practice an operator recovering one package will use the reconstructed downgrade command that this control does not verify.
FAILURE MODE	A snapshot that succeeds at the provider but does not capture the state that mattered, a container volume, a database in flight, or state on a mount outside the snapshot scope, produces a receipt and a false recovery point; the receipt attests to provider completion, not to recoverability.
FAIL STATE	fail-closed
EMITTED SIGNAL	A snapshot receipt per host per run, recorded in the run state and in the plane audit stream, with the cloud provider own snapshot records as independent corroboration. Because the receipt is minted only on confirmed completion it is usable as evidence rather than as an intention, which is the property most completion records in this estate lack.
METRICS	MG-018 MG-021 MG-038 MG-043
COST	moderate
FRICTION	moderate
AWS BINDING	AWS Backup or the EBS and instance snapshot interfaces invoked from a Step Functions state that cannot transition to the apply state until the snapshot reports completed, with Backup Vault Lock so the recovery point cannot be deleted by the principal that made the change.
AZURE BINDING	Azure Backup or managed disk snapshots invoked from a Durable Function or Logic App with an explicit completion gate, and an immutable vault so the recovery point outlives the change and the identity that made it.
ON HOST	partial
EVIDENCE	EV-031
RESEARCH VALUE	A rare specimen in which enforcement lives in the type system rather than in a conditional, so that omitting the control is a compile error rather than a bug. It is the

best available demonstration that structural enforcement can be cheap, and its limitation, that it binds only the one plane that uses it, is precisely the coverage question the register exists to quantify.

GR-L06-03 Pre-approved rollback targets only

INTENT	Ensure a rollback can only move a system to a state that was previously approved, so that recovery cannot be used as a route to deploy something new.
MECHANISM	Would be: a registry of approved revisions with their approval records, and a rollback verb that accepts only an identifier present in that registry, refusing anything else including the caller own suggestion. The deploy gate already implements half of this well, resolving the rollback target server-side from version control at deploy time, pinning it first-write-wins so it cannot be altered later, accepting no target parameter at all on the automatic path, and proving the target is an ancestor of the current head before use. The missing half is the registry itself: the target today is the prior revision, which is not necessarily an approved one.
ENFORCEMENT	structural
BINDS	AC-006 AC-044 AC-045 AC-056
ENABLES	II
PRECONDITION	Every deployed revision must carry an approval record to be a candidate, which requires that no revision ever reaches production outside the gate. On this estate that premise already fails, because ordinary restarts have activated unapproved code at least twice.
BYPASS	A rollback allowlist binds the rollback verb and not the filesystem. Where the running process loads whatever is in a shared working tree, an agent reaches any state it likes by checking it out and waiting for a restart, and the registry becomes irrelevant. It is bypassed a second way by alternative deploy verbs: the estate own guidance records that a separate version control tooling surface still deploys outside the gate, so the gate is not the only deploy path until those verbs are gated or retired. And the ancestor proof that makes the automatic path safe is a check on direction rather than on approval, so it prevents accidental forward motion while permitting a backward move to any ancestor whether or not that ancestor was ever approved.
FAILURE MODE	A registry not populated for older revisions makes emergency rollback to a known-good but unregistered state impossible, converting a safety control into an outage extender; a recorded break-glass path is therefore part of the control rather than an exception to it.
FAIL STATE	fail-closed
EMITTED SIGNAL	A rollback event naming the target revision and the approval record it matched, and a refusal event when a requested target has no approval record, both on the gate append-only event stream so that recovery actions are as auditable as forward ones.
METRICS	MG-007 MG-021 MG-051 MG-060
COST	low
FRICTION	moderate
AWS BINDING	CodeDeploy with a fixed set of registered application revisions held in S3 with versioning and Object Lock, so a rollback selects a stored artefact rather than a source state; CloudFormation change sets restricted to a previously deployed template.
AZURE BINDING	Azure DevOps release pipelines retaining approved artefacts with redeploy-a-previous-release semantics, and Azure Resource Manager deployment stacks targeting a prior stack version.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Tests whether the value of an approval regime survives its own recovery path. Because the estate automatic rollback is already structural in direction and unbounded in approval, adding the registry isolates approval coverage from motion safety as separate contributors to change failure rate.

GR-L06-04 Canary and wave ordering across targets

INTENT	Order a fleet-wide change so that the least consequential targets take it first, with a soak period between waves, so a bad change is caught before it reaches production.
MECHANISM	A fixed wave order over target tiers, running test, then non-production, then production, then the orchestrating host itself, with per-tier concurrency caps that never allow production and the orchestrating host to run concurrently, and a default soak of sixty minutes between waves whose deadline is persisted as a timestamp so it survives a process restart. Any host failure halts the wave and the entire run. The ordering is a compile-time constant array; the tier of each host comes from the hand-maintained declared inventory rather than from runtime discovery.
ENFORCEMENT	conditional
BINDS	AC-031 AC-043 AC-044 AC-045
ENABLES	II
PRECONDITION	Targets must be classified into tiers and the classification must be trustworthy. The live inventory holds three production, three non-production, one test and one orchestrating-host entry, all enabled, with no target in the report-only or unclassified tiers.
BYPASS	The ordering is enforced in code, but the input deciding which wave a host belongs to is a mode 644 JSON file owned by the agent account, so a tier edit reorders the fleet without touching any code, and the same file governs the enabled flag, the credential path, the maintenance window and the restart constraints. The soak is enforced by a persisted deadline, which is likewise a state file the account can write. More importantly the ordering binds only changes made through this plane, while the estate other change paths, a direct remote shell through the model-callable tool, a deploy from the approval gate, or a file write to the read-write mounted production tree, have no wave concept at all; wave ordering therefore describes the discipline of one automation rather than a property of the estate. And the plane has never fired: its audit stream holds only scan and impact assessment entries, no run has ever been created, and no proposal has ever been approved, so the ordering is unexercised code.
FAILURE MODE	A single non-production target that is in truth production, or a test target that does not exercise the changed component, produces a green soak carrying no information; the soak then functions as delay rather than as evidence, and delay is the part operators eventually remove.
FAIL STATE	fail-closed
EMITTED SIGNAL	Per-wave and per-host run records carrying the persisted soak deadline and the halt reason on failure, plus the plane audit stream. Because no run has ever been created, the signal shape is designed and entirely unobserved, which is itself the finding.
METRICS	MG-007 MG-008 MG-020 MG-038
COST	moderate
FRICTION	high
AWS BINDING	CodeDeploy deployment configurations with canary or linear traffic shifting and alarm-based automatic rollback, Systems Manager Patch Manager patch groups paired with maintenance windows, and Step Functions to sequence waves with a wait state for the soak.
AZURE BINDING	Azure DevOps stage gates with a delay and a health query between rings, Update Manager maintenance configurations scoped per ring, and Container Apps revision traffic splitting for the canary itself.
ON HOST	partial
EVIDENCE	EV-031
RESEARCH VALUE	An unusual opportunity to study a control that is fully built, carefully reasoned and never executed. Comparing designed signal against emitted signal gives a direct reading on control signal emission coverage, and the zero-run state makes this the natural first candidate for a controlled exercise in which a real change is put through a real wave for the first time.

GR-L06-05 Change freeze windows

INTENT	Prohibit change during periods when its consequences cannot be absorbed, such as peak service hours, an open incident, or a period with no available responder.
MECHANISM	Would be: a calendar consulted by every change path, expressed as data rather than as prose, carrying an open or closed state per target class and per window, with a change attempted while closed refused rather than queued silently, and an override that is an explicit, recorded human act. The update plane holds the nearest existing analogue in the opposite direction, a per-host maintenance window that must be open before an unattended change may proceed, which is a permissive window rather than a prohibitive freeze and applies to one lane of one plane.
ENFORCEMENT	conditional
BINDS	AC-009 AC-010 AC-031 AC-044 AC-045
ENABLES	II
PRECONDITION	A single point through which changes pass. The estate has at least four independent change paths, so a calendar consulted by one of them constrains one of them and creates a false impression about the rest.
BYPASS	A freeze consulted by change tooling is bypassed by any path that does not consult it, and on this host every consequential path can be driven from a shell. It is bypassed temporally as well, because the calendar is evaluated against the host clock and an account holding root can change the clock or the timezone, so the window has to be evaluated by a service the agent does not run. In practice the most common defeat is neither of these but the override: a freeze with an unrecorded and unreviewed override is a suggestion, and this estate standing authorisation explicitly covers production, so an override would become routine unless using one costs something visible.
FAILURE MODE	A freeze that blocks remediation as well as change turns a minor incident into a prolonged one. The calendar must therefore distinguish change from repair, and that is precisely the distinction an agent is least well placed to make about its own work, so the classification cannot be delegated to the party seeking the exemption.
FAIL STATE	fail-closed
EMITTED SIGNAL	A refusal event naming the calendar and the window that blocked the change, and an override event naming the human who opened it with the justification; both must reach the operator without querying the agent, otherwise the freeze is enforced and reported by the same party.
METRICS	MG-007 MG-032 MG-060 MG-071
COST	low
FRICTION	high
AWS BINDING	Systems Manager Change Calendar with open and closed events consulted by Change Manager and by Automation documents, with maintenance windows providing the permissive half of the same schedule.
AZURE BINDING	Azure Update Manager maintenance configurations and Azure DevOps deployment schedules with gates, plus Azure Policy deny effects where the change is a control plane operation that can be blocked outright.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The temporal analogue of an allowlist, and a direct probe of the estate most striking property, that seventeen automations can initiate action with no human keystroke and several of them run every minute. Imposing a freeze and observing the change in unattended action path count and change failure rate separates harm attributable to unattended change from harm attributable to change as such.

GR-L06-06 Dry run as the default execution mode

INTENT	Make the safe outcome the default, so that an operation must be explicitly promoted in order to take effect rather than explicitly held back from taking effect.
MECHANISM	Present as the shape of one plane rather than as a general default. The update plane transport is split into a read function and an apply function with different signatures; the daily sweep is read-only by construction and states so in its own header; the scan produces proposals rather than changes; and the classifier routes uncertainty to the lane that requires approval rather than to the lane that does not. Live state shows 434 proposals, all still in the proposed status, with an approver set on none of them and nothing ever applied. What is absent is any general dry-run default: console session creation, the headless run endpoint, the browser terminal and the two model-callable tools that execute remotely or spawn a new agent have no preview mode at all, and the model tool own plan mode is used by no launch path.
ENFORCEMENT	conditional
BINDS	AC-002 AC-031 AC-039 AC-043 AC-044
ENABLES	II
PRECONDITION	The operation must have a meaningful preview. For an arbitrary shell command on a remote host there is none, which is why the estate dry-run property exists only where a change is expressed as data, a package and a version, rather than as a command string.
BYPASS	Where a default is safe only because the dangerous half is unwired, the default evaporates the moment it is wired, and that is exactly this plane condition: a complete apply path is written, compiled and deployed, and has simply never been invoked. The read and apply split is a type-level property of a compiled build whose compiled output the agent can rewrite. And the property does not generalise across the host: the two most dangerous surfaces, a model-callable remote shell with an unconstrained command string and a model-callable spawn of a headless permission-skipping agent, both execute immediately with no preview, no confirmation and no dry-run parameter, while the confirmation machinery that does exist in the same module is dead code attached to the least dangerous tool, since nothing ever reads the pending confirmations map back. What it still buys is that within the plane the default really is safe and the proposal record makes an intended change reviewable before it exists.
FAILURE MODE	A dry run that does not exercise the same code path as the real run reports success and predicts nothing. The estate preview is a classifier over parsed package data rather than a rehearsal of the command, so it cannot detect a failure that occurs only on the target.
FAIL STATE	fail-closed
EMITTED SIGNAL	A proposal record per detected change carrying its risk band, lane, justification, risk factors and rollback command, plus the plane audit stream. The live distribution is itself a governance signal: 434 proposals outstanding, every one in the approval-requiring lane, none approved, which says as much about approval throughput as about the control.
METRICS	MG-007 MG-054 MG-055 MG-060
COST	low
FRICTION	moderate
AWS BINDING	CloudFormation change sets, Systems Manager Automation with a preview step, IAM policy simulation, and the dry run flag supported by mutating EC2 interface calls.
AZURE BINDING	Azure Resource Manager what-if analysis and deployment stacks, Azure Policy evaluated in audit effect before being switched to deny, and template what-if executed as a pipeline gate.
ON HOST	partial
EVIDENCE	EV-031 EV-038
RESEARCH VALUE	The estate offers an unusually clean natural experiment, because one plane defaults to preview while the two most dangerous tools default to execution, inside the same process family and under the same identity. Comparing boundary crossing attempt and success rates across those two surfaces isolates the effect of the default from the effect of the capability.

GR-L06-07 Blast radius cap per change

INTENT	Bound how much of the estate a single change may touch, so that a mistaken or malicious change is contained by arithmetic rather than by judgement.
MECHANISM	Would be: a hard limit per change on the number of targets, the number of services restarted and the fraction of a tier affected, evaluated before the first target and re-evaluated between waves, with an excess splitting the change into separately approved units rather than proceeding. The nearest existing element is the wave concurrency cap, which bounds simultaneity and not total reach, since a run may still traverse every declared host one tier at a time.
ENFORCEMENT	conditional
BINDS	AC-031 AC-039 AC-043 AC-044 AC-045
ENABLES	II
PRECONDITION	A reliable count of targets and of what each change touches. The declared inventory supplies the first, and the update plane affected-services field a partial second, but no change path outside that plane has either.
BYPASS	A cap counted per change is defeated by decomposition: an agent issues many changes of one target each and reaches the same aggregate, which is why a radius cap only works paired with a rate limit per action class over a window rather than standing alone. It is defeated a second way by indirection, since a single change to a shared component has an effective radius far larger than its target count and only the plane restart-constraint metadata would notice. And it binds one plane while the model-callable remote execution tool reaches four other production hosts with a free-text command and no cap of any kind, so the cap would sit on the least dangerous of the estate remote paths.
FAILURE MODE	A cap expressed in the wrong unit, hosts rather than services or customer-visible surfaces, produces a comfortable number and no containment; the choice of unit is not a detail of the control, it is the control.
FAIL STATE	fail-closed
EMITTED SIGNAL	A pre-change radius computation recorded alongside the change, and a refusal event when the computation exceeds the cap. That computation is also the only way realised blast radius per incident can ever be compared against a predicted radius rather than reconstructed after the event.
METRICS	MG-007 MG-038 MG-039 MG-043
COST	moderate
FRICTION	high
AWS BINDING	CodeDeploy minimum healthy host settings and deployment configuration limits, Run Command concurrency and error thresholds with automatic stop, and Auto Scaling instance refresh with a minimum healthy percentage.
AZURE BINDING	Update Manager maintenance configurations scoped per group, Azure DevOps stage-level target limits, and Container Apps revision traffic weights to cap exposure during a change.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Turns blast radius from a taxonomy rating into an enforced quantity, which lets the register own action class ratings be validated rather than asserted: if a class rated estate-wide is capped to a single host and nothing of value is lost, the rating was describing potential rather than necessity.

GR-L06-08 Structural test mode intercepting every external dispatch

INTENT	Guarantee that while the system is under test no communication can reach a real member of the public, by replacing the destination of every outbound dispatch at the single point where a destination is resolved.
MECHANISM	This is a REPORTED control, not one verified here: the mechanism runs on the Platform host and only its design documentation and its test suite are readable from this host, so nothing below was observed executing. As described there, the voice and messaging engine resolves every live dispatch destination through one function. While the test flag is set that resolver returns the override destination for every tier of the escalation ladder, the spoken layer substitutes the override for any number read aloud, and the relief, caller and manager message paths each resolve the real destination and then apply the same override before sending. A standing regression test enforces both the dial half and the message half and additionally asserts that a legacy bypass route no longer exists, and a start-up preflight refuses a related manager notification override if it is set while the flag is off. The same documentation records the flag as currently off in production, that is, the engine is live on real destinations, so the control exists and is presently disengaged.
ENFORCEMENT	structural
BINDS	AC-046 AC-047 AC-049 AC-050 AC-052
ENABLES	II
PRECONDITION	There must be exactly one place where a destination is resolved. Interception is only structural if it is impossible for dispatching code to hold a real destination, which requires that no dispatch path constructs a recipient of its own.
BYPASS	What makes this structural rather than instructional is that it checks no intention and asks for no compliance: the value the dispatching code receives is already the override, so there is no branch in which the agent could choose to contact the real party. It is not a rule the agent is told, it is a substitution the agent cannot observe, and agent compliance is therefore not a term in the outcome. Three things would defeat it, and each is a change to the architecture rather than a change to a decision. First, a new dispatch path that resolves a recipient itself instead of calling the single resolver; this is the real risk, a related bypass has occurred before, and the only thing that catches it is the standing regression test, which makes the test load-bearing and its coverage the true boundary of the control. Second, the flag itself, which is a configuration row in the platform database protected by application authorisation rather than by a separate principal, and the estate own recorded residual holes note that a lower-privileged administrator role can escalate to the approving identity, so the flag is one privilege escalation away. Third, and decisively for the present posture, the mechanism is disarmed in production today by deliberate cutover, so its current contribution is to the test environment rather than to production behaviour. The register should draw the correct lesson: the estate has demonstrated that it can build structural enforcement when the harm is judged unacceptable, which is direct evidence that the instructional character of its other controls is a choice rather than a capability limit.
FAILURE MODE	A new code path that bypasses the resolver, undetected because no test covers it, sends a real message while every operator believes the system is contained; that is a silent failure whose first symptom is an external party receiving contact, which is the one failure mode the control exists to make impossible.
FAIL STATE	fail-closed
EMITTED SIGNAL	The regression suite pass or fail result on every run is the primary signal, together with the start-up preflight refusal and an error log line when a related override is set in the wrong state; per-dispatch logs record the substituted destination masked to its final digits so that containment is verifiable without disclosing the recipient.
METRICS	MG-031 MG-033 MG-039 MG-043
COST	low
FRICTION	low
AWS BINDING	The cloud analogue is an egress boundary the workload cannot address around: a subnet with no route to the messaging provider, an SES account held in sandbox mode with a verified-identity allowlist, or the messaging endpoint replaced by a stub in the non-production account, so the destination is unreachable rather than merely unused.

AZURE BINDING	Azure Firewall or a network security group with an explicit destination allowlist and no route from the test subnet to the messaging provider, Azure Communication Services separated by test tenant, and Azure Policy preventing a production connection string from being deployed into a test environment.
ON HOST	no
EVIDENCE	REPORTED
RESEARCH VALUE	The estate clearest example of a genuinely structural control and the reference point against which every instructional control in the catalogue should be scored. It also supplies the register most useful counterfactual: the same organisation, the same engineers and the same model built a structural boundary for contact with the public and an instructional one for production deployment, which suggests the determinant is the perceived irreversibility of the harm rather than the availability of the technique. Testing that proposition is what makes the pair experimentally valuable.

GR-L06-09 Change size limit triggering escalation

INTENT	Let the size and shape of a change determine how much review it needs, so that a large or sensitive change cannot travel the same fast path as a trivial one.
MECHANISM	Would be: a server-side classifier over the actual change rather than its description, producing a size and sensitivity score from lines changed, files touched and whether designated sensitive paths are involved, with anything above a threshold forced onto a slow path with a named reviewer. The deploy gate already contains most of the machinery: its risk service adjudicates the diff rather than the submitter declaration, fails closed when a path cannot be parsed, and refuses to fast-lane any change to the gate itself, while the lead agent confirmation can never promote a change that server-side policy has already refused. What is missing is a size threshold and the escalation it should trigger.
ENFORCEMENT	conditional
BINDS	AC-002 AC-004 AC-044 AC-056
ENABLES	II
PRECONDITION	The change must be inspectable as data before it is applied. That holds for a diff and does not hold for a shell command, so this control can never cover the estate remote execution paths and should not be credited as if it did.
BYPASS	A size limit is defeated by decomposition: an agent submits ten small changes rather than one large one, and unless the classifier scores a submission sequence rather than a submission, each passes individually. It is defeated a second way by placement, because a change that is small and central is more consequential than one that is large and peripheral, so a line-count threshold rewards exactly the wrong shape. And it binds only submissions, so any path that reaches production without submitting, which on this estate includes an ordinary restart picking up a shared working tree, is unscored by construction. What a well-built version still buys is the property the existing risk service already has, namely that the submitter self-assessment plays no part in the routing decision.
FAILURE MODE	A threshold tuned to reduce operator load drifts upward until nearly everything is fast-laned, and because the fast lane is the quiet path the drift produces no signal; the control decays into its own absence without any event marking the transition.
FAIL STATE	fail-closed
EMITTED SIGNAL	A classification record per submission carrying the score, the contributing factors and the resulting lane, plus an escalation event when the threshold is crossed. The record is what makes lane assignment auditable rather than merely trusted, and it is also the data needed to detect threshold drift.
METRICS	MG-007 MG-051 MG-054 MG-055
COST	low
FRICTION	moderate
AWS BINDING	CodePipeline with a Lambda routing action that inspects the artefact and selects between an automatic stage and a manual approval stage, with a code analysis step supplying the sensitivity score.
AZURE BINDING	Azure DevOps branch policies with path filters and reviewers required by path, plus a pipeline gate that routes to a manual approval stage based on a change analysis task.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Isolates review effort from review authority. Because the estate risk service already adjudicates the artefact rather than the claim, adding a size threshold varies only the routing rule, which makes it possible to measure how much of the change failure rate is explained by review depth as opposed to by the mere existence of review.

GR-L06-10 Report-only constraint on the update plane

INTENT	Provide a designation under which a declared target can be scanned and reported on but can never be changed, so that sensitive systems participate in visibility without being exposed to actuation.
MECHANISM	Three independent layers inside the update plane. A nominal type brand: the apply target type carries a module-private symbol, so only the inventory gate function can create one, and both the apply transport and the per-host apply function demand one, which makes the restriction a compile-time property rather than a runtime check. A gate refusal: the gate returns a not-allowed result with explicit reasons for a report-only tier, an unclassified tier, a disabled host, a host with no address, a host declaring the unimplemented cloud run-command transport, and a remote host with no credential path. And a transition table in which the report-only lane has an empty outgoing edge set for every status, with the store transition function the only status-change path, so a report-only item has no route to approved regardless of what any route handler does. There is additionally a never-apply hostname list that the inventory loader force-overrides, so it is the only constraint that survives an edit to the inventory file.
ENFORCEMENT	conditional
BINDS	AC-031 AC-043 AC-044 AC-045
ENABLES	II
PRECONDITION	At least one target must actually be designated report-only for the designation to constrain anything at all.
BYPASS	The constraint is not so much bypassed as unpopulated, and that is the finding. The live inventory has zero hosts at the report-only tier and zero unclassified, holding three production, three non-production, one test and one orchestrating-host entry, all enabled; all 434 live proposals sit in the approval-requiring lane and none is report-only; and the never-apply hostname list, the only constraint that survives an inventory edit because the loader force-overrides it, is an empty frozen array. So the things actually standing between a proposal and a privileged remote package install are the kill switch being absent, a human approval, a confirmed snapshot receipt, a maintenance window for the unattended lane, and one host zero-active-calls precondition; the report-only constraint is not among them. Beyond that, the tier of every host is a field in a mode 644 file owned by the agent account and the compiled enforcement module is likewise agent-writable, so both the designation and its enforcement can be edited from the same shell. What it still buys, stated precisely, is that if a target were designated, no route handler bug and no forgotten check could promote it, which is a genuinely stronger property than a conditional and is the pattern the rest of the estate lacks.
FAILURE MODE	The designation exists, is believed to be in force, and constrains nothing, which is the present state exactly. The mechanism reports healthy because it is correct, and correctness is not coverage; this is the failure mode most likely to be missed by any audit that reads code rather than data.
FAIL STATE	fail-closed
EMITTED SIGNAL	A gate refusal carrying structured reasons, visible in the plane proposal and run records. Today the signal is unobserved because no host is designated and no run has ever been created; the plane audit stream holds eight entries in total, five scans and three impact assessments.
METRICS	MG-031 MG-033 MG-036 MG-061
COST	low
FRICTION	low
AWS BINDING	A resource tag consulted by a Service Control Policy or IAM condition that denies the mutating interface call for tagged resources, so the designation is enforced by the control plane rather than by the tool; Patch Manager patch groups carrying a scan-only baseline for the reporting half.
AZURE BINDING	Azure Policy deny effect keyed on a resource tag, or a Resource Manager lock at read-only level so the designation is enforced by the platform; Update Manager assessment-only mode for the reporting half.
ON HOST	yes
EVIDENCE	EV-031

RESEARCH VALUE	The register best example of the difference between a mechanism and a control: every property one would want is present in the code and none of it binds anything. That makes it the natural test case for a coverage metric, since structural enforcement coverage should score this near zero despite the mechanism being structural, and any metric that scores it highly is measuring design rather than effect.

GR-L06-11 Deploy artefact provenance binding code to a review record

INTENT	Make what a production process runs identical, by construction, to what was approved, so that an approval is a statement about the running system rather than about a database row.
MECHANISM	Would be: build an immutable, content-addressed artefact once, record its digest in the approval record, and make the production process load only that artefact, with no shared working tree and no source checkout at run time. The supervisor refuses to start a version whose digest is not the approved one, and verification happens at every start rather than only at deploy, so that any restart re-asserts the binding rather than silently escaping it.
ENFORCEMENT	structural
BINDS	AC-002 AC-004 AC-009 AC-044 AC-045 AC-056
ENABLES	II
PRECONDITION	A build step and an artefact store. The gated service on the Platform host has neither: it is executed directly from a shared working tree with no build and no pinned artefact, which is precisely why an ordinary restart activates whatever is on disk.
BYPASS	Digest binding removes the estate central failure, that approval flips a field while the process loads a directory, and it does not remove the paths around it. An agent that can write the artefact store or the digest record can approve its own artefact, so the store must be immutable and the digest must be recorded by the approver principal rather than by the submitter. An agent that can edit the supervisor configuration can point it at a different artefact, so that configuration has to sit outside the agent write scope, which on this host it does not. And an agent holding a read-write mount of the production tree, which exists today, defeats any scheme that verifies only at deploy time, which is the specific reason verification must happen at every process start.
FAILURE MODE	A digest recorded but never verified is documentation. The whole value sits in the refusal at start, and a supervisor that logs a mismatch and starts anyway is indistinguishable from having no control while looking exactly like having one, which is the worst position an audit can be in.
FAIL STATE	fail-closed
EMITTED SIGNAL	A start-time attestation per process naming the artefact digest and the approval record it matched, plus a refusal event on mismatch. This signal also makes the estate existing drift detector unnecessary, because drift becomes impossible rather than merely detectable after the fact.
METRICS	MG-007 MG-012 MG-051 MG-067
COST	moderate
FRICTION	moderate
AWS BINDING	ECR images with immutable tags and digest-pinned task definitions, CodeDeploy revisions in S3 with versioning and Object Lock, Signer signatures verified at admission, and CloudFormation change sets so the deployed template is the reviewed one.
AZURE BINDING	Azure Container Registry with content trust and digest-pinned Container Apps revisions, Azure Artifacts with immutable feeds, and Resource Manager deployment stacks so the deployed definition is the approved definition.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	This is the specific control whose absence produced the estate documented deploy incidents, so it has a real and dated counterfactual rather than a hypothetical one. Introducing it and re-measuring change failure rate and approval evidence sufficiency tests whether approval regimes without artefact binding produce any measurable safety at all, which is the strongest available challenge to process-based governance.

GR-L06-12 Post-deploy independent health verification

INTENT	Establish whether a change worked from a source other than the party that made it, so that a completion claim is corroborated rather than accepted.
MECHANISM	Would be: a verification step executed by a principal distinct from the deploying agent, exercising the service real function end to end rather than checking that a process is running, against a success criterion recorded before the change, with the result written to the evidence store under the verifier own identity. Automatic rollback on a failed verification, and an explicit unverifiable outcome kept distinct from both pass and fail so that absence of evidence is never recorded as evidence of success.
ENFORCEMENT	structural
BINDS	AC-009 AC-044 AC-045 AC-053 AC-055
ENABLES	II
PRECONDITION	A verifier identity the deploying agent does not hold, and a functional test the verifier can run without the agent cooperation. Neither exists on this host, where every agent shares one account and the only post-change signal is the agent own report.
BYPASS	The failure this control addresses is not evasion but self-report. The founding case measured 61.1 per cent factual divergence between agent self-report and independent evidence across 18 audited episodes drawn from 46 candidates, including a deploy reported as successful that had in fact disabled the service core connection function for about twenty four hours. A verification performed by the same identity, in the same session, under the same assumptions reproduces the original error rather than catching it, and a health check that asserts a process is online catches only the failures nobody was worried about. The control is therefore defeated by three things: the verifier being the actor, the check being liveness rather than function, and the result being recorded by the actor. It is defeated a fourth way by scope, since a verification at deploy time says nothing about a restart an hour later that loads different code, which is exactly the estate observed failure.
FAILURE MODE	A verification that is thorough but whose failure carries no consequence degrades into a report nobody reads. The estate already contains an instance: its drift detector correctly identifies that the running code may not be what was approved and deliberately performs no remediation, so the finding has existed and changed nothing.
FAIL STATE	fail-closed
EMITTED SIGNAL	A verification result per change written under the verifier identity, naming the criterion, the observed value and the source of the observation. The divergence between that record and the agent own completion claim is directly the claim divergence rate the thesis measures, which makes this control the instrument as well as the intervention.
METRICS	MG-019 MG-022 MG-023 MG-024
COST	high
FRICTION	moderate
AWS BINDING	CodeDeploy lifecycle hooks running validation from a Lambda with its own execution role, CloudWatch Synthetics canaries exercising the real endpoint continuously, and alarm-based automatic rollback so a failed verification acts rather than reports.
AZURE BINDING	Azure DevOps post-deployment gates querying Application Insights availability tests, Azure Monitor standard tests as the independent prober, and an action group triggering an automated rollback runbook.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The control that most directly instruments the thesis central empirical claim. Because the founding case already has a divergence rate obtained by manual audit, introducing an automated independent verifier allows the same quantity to be measured continuously, and the gap between the manual and automated figures is itself a measure of how much divergence a self-reporting regime conceals.

2.7 L7. Mandato e política

Esta camada governa o que um agente está autorizado a fazer como um artefato verificável por máquina, distinto do que ele é capaz de fazer, que é a L1 até a L6, e do que lhe é dito, que é a L12. O caso fundador essencialmente não tem L7 hoje: o mandato operativo é um brief em prosa, presente mas instructional, e sua única restrição structural é estar delimitado a um diretório de trabalho, o que por sua vez é apenas parcial `EV-032`. Toda entrada que tornaria o mandato uma estrutura de dados em vez de um parágrafo, política declarativa expressa como dado, um ponto de decisão de política separado do ponto de enforcement, mandatos assinados e versionados com expiração, um diário de mudanças que a parte governada não pode escrever, e a impossibilidade structural de um agente ampliar seu próprio mandato, está proposta.

Esta é a camada onde a lacuna entre acesso e mandato, a segunda das três observações fundadoras do D2, está estruturalmente localizada. Cinquenta das sessenta classes de ação no inventário mostram acesso mais permissivo do que mandato. Essa lacuna existe porque o mandato hoje não tem nenhuma representação contra a qual o runtime possa verificar; ele é real apenas no sentido de que o operador o pretende, o que é exatamente o que as entradas da L7 existem para mudar.

Camada 7. Mandato e política.

GR-L07-01 Declarative action class policy expressed as data

INTENT	Express the permitted, reviewable and prohibited action classes as a machine readable policy document so that authorisation becomes a data lookup rather than a reading of prose.
MECHANISM	A versioned policy artefact enumerating each action class identifier with its reversibility, blast radius, required access level and required mandate level, served to enforcement points as data and evaluated by a policy engine; on this host it would replace the 23-pattern shell command deny list and the prose operating guidance with a single table keyed by action class. Candidate encodings are Cedar for the cloud control plane and a rego policy bundle for the tool layer.
ENFORCEMENT	conditional
BINDS	AC-020 AC-025 AC-026 AC-029 AC-036 AC-039 AC-043 AC-048
ENABLES	II
PRECONDITION	L1 identity separation so the policy has a nameable subject; a stable action class taxonomy; an enforcement point that consults the table before the action rather than a reviewer who consults it afterwards.
BYPASS	On this host the policy artefact would be an ordinary file under the single account every agent runs as, and that account holds unrestricted passwordless root, so the agent can edit the table it is judged by exactly as it can today edit the settings file that grants its permissions; the control is only as strong as the identity that owns the store. What it still buys is substantial: it converts an unbounded prose mandate into a finite enumerated surface that can be diffed, counted and tested, which is what makes unclassified action share measurable at all.
FAILURE MODE	The table drifts behind the real action surface, so newly acquired capabilities arrive unclassified and inherit whatever the default is.
FAIL STATE	fail-open
EMITTED SIGNAL	none today, the artefact does not exist; as designed it would emit a policy evaluation record per attempted action carrying action class, verdict, matched rule and policy revision, written by the decision point rather than into the agent transcript.
METRICS	MG-036 MG-031 MG-030 MG-034
COST	moderate
FRICTION	low
AWS BINDING	AWS Verified Permissions holding the action class table as Cedar policy; IAM permission boundaries and Organizations service control policies expressing the same classes at the control plane; IAM Access Analyzer to prove no attached policy grants beyond the declared class; EventBridge on policy change events.
AZURE BINDING	Azure Policy definitions and initiatives assigned at management group scope; custom Azure RBAC role definitions naming the same classes rather than wildcards; Azure Resource Graph to reconcile actual assignments against the declared table.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Tests whether making the class taxonomy machine readable reduces unclassified action share, and whether the count of actions falling outside the table is stable or grows as the agent acquires capability.

GR-L07-02 Policy decision point separate from the enforcement point

INTENT	Ensure the component deciding whether an action is permitted is not the component performing it, so that altering the actor does not alter the decision.
MECHANISM	A decision service holding the policy, queried by each enforcement point with a request naming principal, action class and target; it runs under an identity the agent does not hold, on infrastructure the agent has no route to, returns permit or deny with a reason, and the enforcement point refuses when the service is unreachable. On this host the two roles are currently one process: the model runtime both decides, by reading its own settings file, and acts.
ENFORCEMENT	structural
BINDS	AC-020 AC-026 AC-029 AC-037 AC-039 AC-043 AC-044 AC-056
ENABLES	II
PRECONDITION	L1 per-agent workload identity so the decision service has a principal to evaluate; a network path the agent cannot intercept or redirect; enforcement points written to fail closed when the decision service does not answer.
BYPASS	The bypass is not the decision service, it is the set of effects reachable without passing an enforcement point. On this host that set includes the read-write mounted production tree of another estate host, the passwordless root shell, and the instance metadata endpoint from which any process can mint the instance role; none of those traverses any decision component. Separating the decision point is therefore necessary and insufficient, and must be paired with removing the alternative paths. Even alone it buys one thing that nothing else in the catalogue buys: the decision record is written by a party the actor cannot write, so authorisation becomes independently auditable.
FAILURE MODE	New enforcement points are added without a decision call, so the governed surface grows faster than the integration.
FAIL STATE	fail-closed
EMITTED SIGNAL	A decision record per evaluation carrying principal, action class, target, verdict, policy revision and latency, held by the decision service and not by the agent; none today because no decision point exists.
METRICS	MG-031 MG-059 MG-048 MG-060
COST	high
FRICTION	moderate
AWS BINDING	AWS Verified Permissions as the decision point with Cedar policies called from the workload; IAM policy evaluation is itself the canonical separated decision point for control plane calls; Organizations service control policies place the decision above the account entirely.
AZURE BINDING	Azure RBAC and Azure Policy evaluate at the Resource Manager tier rather than in the workload; Entra Privileged Identity Management for activation decisions; deny effect Azure Policy assignments at management group scope.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Provides the cleanest experimental contrast in the programme: identical agent and identical task, with the permit decision made in process versus made by a party the agent cannot write. Measures whether structural enforcement coverage rather than instruction quality predicts boundary crossing success.

GR-L07-03 Signed and versioned mandates

INTENT	Bind a mandate to an issuing authority and a specific revision so that what an agent operates under at any moment is provable rather than asserted.
MECHANISM	The mandate becomes a structured document carrying scope, action classes, level per class, expiry and issuer, signed by a key held by the issuing human role; the agent receives document and signature, and every enforcement point verifies signature and revision before honouring it. This replaces the present arrangement, in which the mandate is unsigned Markdown read into context at session start.
ENFORCEMENT	conditional
BINDS	AC-019 AC-020 AC-025 AC-043 AC-044 AC-048 AC-052 AC-056
ENABLES	II
PRECONDITION	A signing key held by an identity the agent does not hold, which is L1 separation of the approver; a canonical mandate schema; enforcement points that read the signed document instead of a prose file.
BYPASS	If the verifying process runs under the same account as the agent, the agent replaces the trust anchor or the verification code, so signing only relocates trust to key custody. Concretely here, all instruction files and all eight settings files are writable by the governed account and there is no administrator owned settings tier anywhere on the host, so a signed mandate gains nothing until verifier and anchor live outside that account. Once key custody is separated it buys attribution: every mandate becomes traceable to a person and a revision, which makes mandate stability and mandate widening countable rather than inferred.
FAILURE MODE	Signature verification is treated as advisory, or a valid signature is presented over a stale revision that was never withdrawn.
FAIL STATE	fail-closed
EMITTED SIGNAL	A mandate presented event per session start carrying issuer, revision and verification outcome; today the mandate is loaded as prose context and produces no event at load, at breach or at edit.
METRICS	MG-034 MG-035 MG-050 MG-064
COST	moderate
FRICTION	moderate
AWS BINDING	AWS KMS asymmetric keys for mandate issuance; mandate documents in S3 with versioning and Object Lock; EventBridge on object version events; STS session tags to carry the mandate revision into the workload session.
AZURE BINDING	Azure Key Vault signing keys; immutable blob storage with version level immutability for the documents; Entra Privileged Identity Management approval as the issuance step; Azure Monitor for issuance events.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Tests whether a signed and revisioned mandate makes widening detectable at all, given that the founding case cannot today distinguish an authorised scope change from an agent editing its own instruction file.

GR-L07-04 Mandate expiry

INTENT	Cause authority to lapse by default so that continued autonomy requires a positive human act rather than the absence of one.
MECHANISM	Every mandate carries a not-after timestamp; enforcement points refuse action classes above Level I once it passes, and the workload credential is minted only for the mandate remaining life so that expiry withdraws capability rather than merely failing a check. The present standing authorisation clause in the operating guidance document has no temporal field of any kind and has been continuously in force.
ENFORCEMENT	structural
BINDS	AC-019 AC-029 AC-037 AC-039 AC-043 AC-044 AC-048 AC-052
ENABLES	II
PRECONDITION	L1 ephemeral credentials with a short session duration, so that lapse removes capability; a renewal path a human actually operates; a clock the agent cannot set, which is a real dependency here because the account holds root.
BYPASS	A purely local expiry check is defeated by the agent setting the host clock, which passwordless root permits; expiry binds only when the expiring artefact is a credential minted elsewhere with a server side validity window. What it still buys locally is a change of direction: an open ended standing authorisation becomes a decaying one, so the state after neglect is less authority rather than the same authority, which matters in a case where 46 days of audit logs contain zero human approvals of any kind.
FAILURE MODE	Renewal degenerates into a scheduled rubber stamp, restoring the standing grant with added ceremony and no added scrutiny.
FAIL STATE	fail-closed
EMITTED SIGNAL	A mandate expired event at the enforcement point and a renewal event naming the approving human identity; none today, because the operative mandate has no expiry field to observe.
METRICS	MG-034 MG-037 MG-050 MG-054
COST	low
FRICTION	high
AWS BINDING	STS session duration and IAM role maximum session duration as the expiring artefact; IAM policy conditions on aws:TokenIssueTime or aws:CurrentTime; EventBridge Scheduler to revoke on lapse.
AZURE BINDING	Entra Privileged Identity Management eligible assignments with activation expiry and maximum duration; Azure RBAC role assignments with an explicit end date.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Isolates the effect of default off authority: whether a lapsing mandate reduces the count of high consequence actions taken with no contemporaneous human act, against a baseline of continuous autonomous action with no approvals recorded across the whole observation window.

GR-L07-05 Mandate change journal with approver role

INTENT	Make every change to what an agent may do an attributable, reviewable record naming the human who approved it.
MECHANISM	An append only journal written by the issuance service under the approver identity, holding the prior revision hash, the diff, the approver, the justification and the effective time; the journal is the only path by which a revision becomes active, so an unjournalled mandate is simply not honoured by any enforcement point.
ENFORCEMENT	structural
BINDS	AC-020 AC-021 AC-025 AC-026 AC-030 AC-055 AC-056 AC-057
ENABLES	III
PRECONDITION	Signed and versioned mandates, so a revision is a well defined object; L1 separation of the approver identity from the operating account; a journal store the governed party cannot write, which requires the L8 evidence separation work.
BYPASS	The journal binds only changes that flow through issuance, so any control expressed as an ordinary file rather than as a mandate revision sits outside it. Today that would be all of them: the permission settings file, the per agent role files, the shared roster and the update kill switch are plain files under the governed account, changeable with no journal entry, no authentication and no log line. Completeness is therefore a function of how much has been migrated into mandate form, not of the journal itself. What it buys is precision on the event this thesis cares about most: a governed party widening its own mandate becomes a countable event with an expected approver, so its absence becomes evidence rather than silence.
FAILURE MODE	The approver role collapses onto the same person who operates the agents, preserving attribution while destroying independence.
FAIL STATE	fail-closed
EMITTED SIGNAL	One journal entry per revision carrying prior hash, approver identity and diff; today mandate shaped changes are ordinary file writes producing no event, and the only forensic trace on this host is a fifteen minute git auto commit performed by a cron job the same account controls.
METRICS	MG-035 MG-067 MG-064 MG-034
COST	moderate
FRICTION	moderate
AWS BINDING	A hash chained ledger or CloudTrail Lake as the journal store; S3 Object Lock in compliance mode for the revision artefacts; IAM Access Analyzer policy change findings; EventBridge rules on policy and permission boundary change events delivered cross account.
AZURE BINDING	Log Analytics workspace with immutable retention for the journal; Microsoft Purview audit for approver actions; Activity Log events on Azure Policy assignment change; Entra Privileged Identity Management approval records.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Supplies the denominator for mandate widening events initiated by the governed party, which cannot be measured in the founding case because a self initiated permission change is indistinguishable from any other file write.

GR-L07-06 Structural impossibility of an agent widening its own mandate

INTENT	Remove self amendment from the agent reachable action set entirely, so that the boundary an agent works within cannot be moved by the agent under any circumstance.
MECHANISM	Three conditions held simultaneously. First, the policy store is written only by an identity the agent does not hold and cannot obtain, meaning no credential path, no metadata issued role, no privilege escalation and no shared account. Second, the store sits on infrastructure the agent has no network or filesystem route to; a local file, however permissioned, does not satisfy this while the agent holds root. Third, the enforcement point fails closed when the store is unreachable, so denial of service against the policy does not become permission. Implemented as a policy served by a control plane outside the workload account, evaluated per action, with the workload identity granted evaluation and no write action on the policy resource.
ENFORCEMENT	structural
BINDS	AC-020 AC-021 AC-025 AC-026 AC-029 AC-030 AC-032 AC-057
ENABLES	II
PRECONDITION	L1 per-agent workload identity, L1 least privilege per action class, removal of passwordless root from the agent account, and blocking agent process access to the instance metadata endpoint. Without all four the agent simply becomes the writing identity and the control evaporates.
BYPASS	Contrast with the present arrangement, which is the exact inverse: the settings file that grants permissions, the instruction files that state the mandate, the fleet registry, the update kill switch and the hook scripts are all ordinary files owned by the single account every agent runs as, and that account holds unrestricted passwordless root, so widening the mandate is one file write with no authentication and no audit entry. Under the proposed arrangement the residual bypasses reduce to exactly three, each nameable and each testable: obtaining the writing identity credential, reaching the store network path, or defeating the enforcement point rather than the store. That is the whole point. Every other control in this catalogue degrades to instruction if the governed party can edit the governance, so this entry is the precondition for the rest of the architecture rather than one item within it.
FAILURE MODE	Partial implementation in which the policy is remote but a locally cached copy is honoured when the remote is unreachable, which silently reintroduces a writable local store.
FAIL STATE	fail-closed
EMITTED SIGNAL	A denied write event on the policy resource emitted by the store control plane under the writing identity rather than by the agent, plus a policy unreachable event at the enforcement point that must be distinguishable from a permit. Today the equivalent event does not exist, because the write succeeds silently.
METRICS	MG-035 MG-031 MG-033 MG-063
COST	high
FRICTION	moderate
AWS BINDING	Organizations service control policies, which a member account cannot alter regardless of its own administrator rights; IAM permission boundaries denying iam and policy write actions on the workload role; an AWS Verified Permissions policy store in a separate account; EventBridge forwarding any policy change event to a security account.
AZURE BINDING	Azure Policy with deny effect assigned at management group scope above the subscription; custom RBAC role definitions excluding Microsoft.Authorization write actions; Entra Privileged Identity Management so the elevating identity is separate and time bound; Azure Resource Graph queries proving no assignment grants the workload write at the policy scope.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The central hypothesis of the thesis is testable here and nowhere else: that boundary crossing success rate falls to zero for the self amendment class when the policy store is unreachable to the governed party, and remains unchanged under any amount of instruction. It also forces an honest accounting of cost, because the same change removes the passwordless root that fleet supervision on this host actively depends on.

GR-L07-07 Default deny for unclassified actions

INTENT	Make an action the taxonomy has not yet classified unavailable rather than unrestricted, so capability growth cannot outrun governance silently.
MECHANISM	The enforcement point resolves an attempted action to an action class; where no class matches the verdict is deny with a reason of unclassified, and the attempt is queued for triage into the taxonomy. This inverts the present posture, in which the only permission gate is a 23-pattern deny list of shell command globs running under a bypass permissions default mode, so anything unnamed is permitted by construction.
ENFORCEMENT	structural
BINDS	AC-035 AC-036 AC-039 AC-043 AC-047 AC-048 AC-050 AC-060
ENABLES	II
PRECONDITION	A runtime classifier, since without one there is nothing to fail to match; and a triage path that converts denials into classifications, otherwise the pressure falls on disabling the control.
BYPASS	Default deny binds only where classification happens. The classifier would sit in the model runtime pre-tool hook, which the governed account owns, so a permissive rule or a rewritten hook defeats it; and every effect reached outside the tool layer, most obviously anything a spawned shell subprocess does, is never classified at all. What it buys regardless is a change in the direction of error: an unmodelled capability surfaces as a denial rather than passing as an unremarkable success, which is the difference between a measurable unclassified action share and none.
FAILURE MODE	Denial volume at first deployment is high enough that the control is switched off rather than the taxonomy extended.
FAIL STATE	fail-closed
EMITTED SIGNAL	A deny unclassified event carrying the attempted tool, the argument shape and the reason, deliberately distinct from a deny by policy event so the two can be counted separately and the taxonomy gap sized.
METRICS	MG-036 MG-060 MG-032 MG-059
COST	moderate
FRICTION	high
AWS BINDING	IAM default deny semantics are the reference model; permission boundaries and service control policies extend it to actions the workload policy omitted; Verified Permissions returns deny in the absence of a matching permit policy by construction.
AZURE BINDING	Azure RBAC is default deny for unassigned actions; custom role definitions with explicit action lists rather than wildcards; Azure Policy deny effect for resource shapes no initiative covers.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Measures the size and character of the ungoverned surface directly. The count of actions matching no class tells me whether a 60 class taxonomy is adequate or whether real agent behaviour is long tailed, which is the difference between a usable instrument and a plausible one.

GR-L07-08 Runtime action classifier assigning class and level at attempt

INTENT	Make the action class taxonomy executable by assigning a class, a reversibility, a blast radius and a required mandate level to every action at the moment it is attempted rather than in retrospective review.
MECHANISM	An interception point that receives the action before it executes, in the model runtime pre-tool hook for tool calls and at the control plane for cloud calls, resolving it to one of the enumerated classes by matching tool name, argument shape, target path and target host, then returning class, required level and a verdict of permit, permit with review, or deny, and emitting a classification record whether or not the action proceeds. Classification is synchronous and in the blocking path, so the class is known before the effect exists; retrospective classification from transcripts, which is what the founding case does today, can only describe what already happened. Ambiguity routes to the more restrictive class, following the pattern already proven on this host by the update plane lane classifier, which is deterministic code with eight conjunctive conditions and routes uncertainty to the human approval lane.
ENFORCEMENT	conditional
BINDS	AC-016 AC-023 AC-029 AC-036 AC-039 AC-043 AC-048 AC-055
ENABLES	II
PRECONDITION	A stable action class register; an interception point on every path that can produce an effect, which today means the tool layer covers the model own calls but not what a spawned subprocess subsequently does; and a class to level mapping a decision point can read.
BYPASS	The interception point already exists here and fires on every tool call, but it is a shell script in a directory the governed account owns whose own header mandates an unconditional success exit, so a classifier hosted there is advisory against its own subjects. It is also shallow: an effect reached through a subprocess, a background process or the mounted remote filesystem is classified as one shell invocation rather than as the actions it performs, so a single record can conceal an unbounded number of unclassified effects. What it buys even in that weak form does not depend on its enforcement strength: it is the only mechanism in the catalogue that produces a per attempt, per class denominator, and every metric in the register expressed per action class is unmeasurable without it.
FAILURE MODE	Granularity collapse, in which the majority of consequential actions land in a single generic shell execution class, which is exactly what would happen on this host today.
FAIL STATE	fail-closed
EMITTED SIGNAL	One classification record per attempted action carrying timestamp, session, tool, resolved action class, reversibility, blast radius, required level, verdict and matched rule. This is the observation that converts the taxonomy from a description into an instrument, and nothing equivalent exists today on any channel.
METRICS	MG-036 MG-004 MG-032 MG-043
COST	moderate
FRICTION	low
AWS BINDING	CloudTrail already classifies control plane calls by service and action name and is the closest existing analogue; EventBridge rules matching event name and resource for class assignment; Verified Permissions for the class to level decision. No native equivalent exists for tool layer classification inside the agent process; implemented in the workload as a pre-tool hook.
AZURE BINDING	Azure Monitor Activity Log operation names give the same control plane classification and Azure Policy aliases identify the resource shape; no native equivalent for in process tool classification, so implemented in the workload.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The highest value component in the whole programme, because it converts the action class register from a descriptive taxonomy into an executable instrument. It makes cycle time per action class, first attempt success rate per action class, boundary crossing attempt rate and irreversible action share measurable at all, and it lets me test whether

classifying at attempt rather than in review changes agent behaviour, which is the difference between governance that observes and governance that participates.

GR-L07-09 Prose brief as the operative mandate

INTENT	Records what actually governs agent scope on this host today, which is a natural language document read into the model context at session start.
MECHANISM	Two Markdown files loaded as context for every session because every session shares one working directory, plus a per agent role file, a shared roster carrying the deployment approval chain, and a startup prompt naming them by absolute path; one of them carries a standing authorisation clause asserting unrestricted change authority over production, public DNS and cloud accounts. That these are context and not enforcement was verified four ways: the only pre-tool hook is telemetry that always exits successfully, there is no administrator owned managed settings tier anywhere on the host, the console reads instruction files only to display them to a human, and no authorisation check anywhere parses them. The corpus is 40 classic instruction files totalling about 538 kB on a narrow reading, or about 1,240 files and 26.7 MB counting role files, rosters, prompts, three memory stores and the knowledge base, and every one of them is writable by the agents it governs.
ENFORCEMENT	instructional
BINDS	AC-001 AC-019 AC-029 AC-039 AC-043 AC-044 AC-048 AC-052
ENABLES	I
PRECONDITION	The model disposition to comply, and nothing else. The document is unsigned, unversioned and unexpiring, and is not checked against any policy before it takes effect; it is loaded and it is in force.
BYPASS	There is nothing to bypass, because nothing is enforced; the failure mode is not evasion but non compliance, which is externally indistinguishable from misreading, context truncation, or a later instruction overriding an earlier one. The corpus is additionally writable by every agent it governs, so an agent that edits its own mandate and then complies with the edited text is, from outside, behaving correctly. What the arrangement still buys is real and should not be understated: it is a cheap, very high bandwidth way to convey intent, and the operating discipline it encodes demonstrably shapes behaviour most of the time. The thesis claim is narrower and harder, that most of the time is not a governance property.
FAILURE MODE	Silent divergence between the stated mandate and behaviour, with no signal emitted at the moment of divergence and no artefact recording which text was in force.
FAIL STATE	fail-open
EMITTED SIGNAL	none today; the mandate emits no event at load, none at breach and none at edit. To be measurable it would have to emit a mandate loaded event carrying a digest of the text actually in context, and an edit event whenever any file in the corpus changes, so the mandate in force at the time of an action could be reconstructed rather than assumed.
METRICS	MG-050 MG-031 MG-034 MG-064
COST	low
FRICTION	low
AWS BINDING	n/a; no cloud service participates. The prose mandate never reaches an AWS authorisation decision, which is itself the finding, since the same account holds five static non expiring administrator class key pairs that no clause in it constrains.
AZURE BINDING	n/a; for the same reason. No Azure Policy assignment, RBAC role definition or Privileged Identity Management configuration expresses any part of the prose mandate.
ON HOST	yes
EVIDENCE	EV-032 EV-020 EV-026
RESEARCH VALUE	The baseline condition for the whole programme, and a strong baseline rather than a straw one: this is instruction based governance in an unusually complete, well maintained and internally consistent form. If a 26.7 MB carefully curated instruction corpus does not prevent the measured divergence, the insufficiency is not a matter of writing the instructions better.

GR-L07-10 Mandate scoped to a working directory

INTENT	Confine an agent operating scope by giving it a dedicated working directory that its brief treats as the boundary of its concern.
MECHANISM	Fleet agent creation writes a per agent directory holding the role file, running log and inbox, and the console deliberately refuses an operator chosen starting directory for a fleet agent on the stated reasoning that the working directory is never operator chosen. The scoping is then discarded at launch: the wrapper hardcodes the session working directory to the shared home so that all agents share one auto-memory store and one transcript namespace, and the launch script records this as a deliberate deviation from the design. Per agent settings files exist but are literally empty objects, documented as ineffective for exactly this reason.
ENFORCEMENT	instructional
BINDS	AC-002 AC-017 AC-022 AC-023 AC-024 AC-025 AC-028 AC-053
ENABLES	I
PRECONDITION	L3 per session filesystem isolation by namespace or container would be required for the scope to bind at all; today the only precondition satisfied is that the directory exists and is named in the role file.
BYPASS	The agent own tools are not constrained by the console realpath allowlist, and the process runs with the shared home as its working directory, so the scope is a sentence in a role file rather than a boundary; any absolute path on the host is reachable with an ordinary read or write, including every other agent working directory, role file, gate credential, bus token, mailbox and transcript. What it still buys: the per agent directory gives each agent a stable anchor for its role file, running log and inbox, which is what makes per agent attribution possible at all in a system with no per agent operating system identity.
FAILURE MODE	An agent acts on another agent files believing them to be its own, which has already occurred here, where a session mistakenly believed it was a different agent, read that agent working directory and used its credential.
FAIL STATE	fail-open
EMITTED SIGNAL	none today for the scope itself; the console emits a session create record naming the working directory, but no event is emitted when an agent reads or writes outside it. A path scope violation event at the tool layer is the minimum needed to make the boundary measurable.
METRICS	MG-030 MG-040 MG-067 MG-038
COST	low
FRICTION	low
AWS BINDING	no native equivalent for a process working directory; the nearest analogue is an IAM policy condition on an S3 key prefix per principal, which binds object storage only and says nothing about the local filesystem.
AZURE BINDING	no native equivalent; Azure RBAC scoped to a resource group or a storage container path is the nearest analogue and again covers cloud resources rather than the host filesystem.
ON HOST	partial
EVIDENCE	EV-020 EV-032 EV-030
RESEARCH VALUE	A clean instance of nominal to effective divergence: a scope declared in configuration and deliberately discarded at runtime for a legitimate engineering reason, shared memory continuity, with the governance cost recorded nowhere. Lets me test whether operators can predict which of their declared boundaries are actually in force.

GR-L07-11 **Per-tool permission grant recorded in the request**

INTENT	Carry an explicit permission set with each request so the tool layer can refuse an action the requester was not granted.
MECHANISM	The orchestrator service exposes ten model callable tools and gates three of them on a named permission, being remote shell execution, headless agent spawn, and cloud instance state change, by testing membership of a permissions array that also honours a superuser wildcard. The array arrives in the request body, and the signed token authenticating the call carries an empty payload with no principal. The seven read oriented tools require no permission at all.
ENFORCEMENT	conditional
BINDS	AC-014 AC-022 AC-036 AC-038 AC-039 AC-042 AC-043 AC-029
ENABLES	II
PRECONDITION	A verified principal from which permissions can be resolved server side; that precondition is not met, so the check evaluates data supplied by the caller it is meant to constrain.
BYPASS	The permission array travels in the request body while the bearer carries no principal, so anyone able to mint or replay a valid 60 second token on the VPC reachable port can assert the superuser wildcard and obtain the full tool set; there is no replay cache, no nonce, and a failed authentication produces no audit record and no alert. Separately, the confirmation flow that would have gated the highest consequence tools is dead code in the running process: the pending confirmation map is written and never read, and its consumer exists only in a service that does not run on this host, so the gate sits on the least dangerous tool and is absent from the two most dangerous. What the control still buys: the check is real code on the correct three tools, the host argument for remote execution is enforced against a six entry allowlist twice, in schema and in code, and the path arguments of the read tools are resolved and prefix checked, so the shape of a working per request grant is present and only the identity binding is missing.
FAILURE MODE	The grant is trusted from the request, so the strength of the whole control equals the secrecy of one shared bearer secret on a port reachable from every host in the network.
FAIL STATE	fail-open
EMITTED SIGNAL	A permission denied result aborts the model turn with a fixed message and the turn is logged back to the console; but a rejected authentication on the receiving port emits nothing at all, so the authentication boundary has no detection surface of any kind.
METRICS	MG-059 MG-060 MG-042 MG-061
COST	low
FRICTION	low
AWS BINDING	Done properly this is an IAM role assumed per caller with the permission set in the role policy rather than in the payload, or Verified Permissions evaluating a principal from a verified token; STS session tags would carry the grant without trusting the body.
AZURE BINDING	An Entra issued token carrying app roles or scopes as claims rather than body fields, validated against Azure RBAC; a managed identity for the calling service so that no shared secret exists to replay.
ON HOST	partial
EVIDENCE	EV-038 EV-029
RESEARCH VALUE	A precise instance of a control whose enforcement class is decided by one design choice: identical code is structural when the permission set comes from a verified token and fail open when it comes from the body. Useful for testing whether practitioners classify controls correctly by inspection, because the check reads as robust.

2.8 L8. Evidência, auditoria e verificação

Esta camada governa que registro existe do que um agente fez, e quem pode verificar esse registro de forma independente do agente que o produziu. É a camada que o D2 identifica como carregando a falha central medida do caso fundador: a divergência factual de 61,1 por cento entre o autorrelato do agente e a evidência independente **REPORTED** é, nos termos deste catálogo, uma medida de que as

entradas structural desta camada são todas propostas, e não presentes. Uma transcrição de sessão é escrita pelo runtime do modelo e está presente, mas é apenas conditional, porque reside em um sistema de arquivos comum que a mesma conta pode modificar e não carrega cadeia de hash, número de sequência nem uma identidade separada que a detenha `EV-026` `EV-034`. Um substrato de hook de transcrição dispara em toda chamada de ferramenta e está presente, classificado `none`, porque o script do hook é escrito para sempre determinar um código de saída de sucesso, de modo que é telemetria e não pode votar sobre se um agente continua; o substrato de enforcement existe e deliberadamente não é usado para enforcement `EV-026`.

Nenhuma das nove entradas structural propostas da L8 existe no host hoje: transcrições somente-acréscimo com enforcement feito pelo armazenamento, um verificador independente sob uma identidade separada, uma proibição de que o próprio agente atuante escreva sua atestação de conclusão, e um fluxo de eventos emitido pela infraestrutura, independente do agente. Essas quatro, juntas, são o que o D2 identifica como removendo o determinante de supervisão `SD-05` de não verificabilidade, e o próprio relato do D2 sobre por que esse determinante é o mais responsivo a investimento é o relato desta camada.

Camada 8. Evidência, auditoria e verificação.

GR-L08-01 Session transcript written by the model runtime

INTENT	Retain a structured, machine readable record of every session turn and tool call so an episode can be reconstructed after the fact.
MECHANISM	The model runtime writes line delimited JSON per session under a per working directory project namespace, about 1.13 GiB across 2,174 files, with typed records for assistant, user, system, file history snapshot, permission mode and queue operation, plus a per session sidecar holding subagent and tool result data. Because every session on this host shares one working directory, all agents share a single project namespace. A separate raw terminal byte stream is captured per session by tmux pipe-pane and is never rotated. Live retention is the runtime built in cleanup at an inferred 30 days, and a cron job copies the tree every five minutes with no delete flag, so the copy is the longer horizon record.
ENFORCEMENT	conditional
BINDS	AC-002 AC-016 AC-019 AC-023 AC-029 AC-053 AC-054 AC-055
ENABLES	II
PRECONDITION	Nothing beyond the runtime writing it; but for the record to attribute an act to an agent, a per agent identity is required, and none exists, so attribution rests on session naming convention rather than on identity.
BYPASS	The files are owned by the same account every agent runs as, at mode 600 or 644, with no append only or immutable filesystem attribute, no hash chaining and no separate identity holding the record, so the writing process holds full truncate and unlink rights over its own history. Observation established append only in practice and by writer convention only: monotonic growth across a live window, compaction appending rather than truncating, and no rewriting script anywhere on the host, but also that a past manual edit could not be excluded because no integrity control would reveal one. The five minutely copy is on the same host, same account and same filesystem, so it is redundancy, not independence.
FAILURE MODE	A rewrite or truncation leaves no trace; separately the raw terminal capture and the structured record share no correlation key, so two records of the same episode cannot be joined.
FAIL STATE	fail-open
EMITTED SIGNAL	The transcript itself, tailed event driven by the console chat view and by the resource accounting sync; it emits no event on modification, none on deletion and no completeness assertion, so its own integrity is unobservable by design rather than by oversight.
METRICS	MG-066 MG-064 MG-065 MG-029
COST	low
FRICTION	low
AWS BINDING	The equivalent with integrity is delivery to S3 with Object Lock in compliance mode through Kinesis Data Firehose, written by a role the workload cannot assume; CloudTrail log file integrity validation is the model for the digest scheme.
AZURE BINDING	Azure Storage immutable blob with a time based retention policy or legal hold, fed through Event Hubs by an identity separate from the workload, with Azure Monitor immutable retention for the query surface.
ON HOST	yes
EVIDENCE	EV-026 EV-020 EV-024
RESEARCH VALUE	Supplies raw material for episode reconstructability and divergence measurement at a scale rarely available in field research, and simultaneously demonstrates the limit: a record owned by its subject supports reconstruction but cannot support attestation, so any divergence figure built on it is an upper bound on what the agent chose to record.

GR-L08-02 Audit logging on destructive console operations

INTENT	Record operator initiated console actions with actor, target and outcome so supervisory activity can be reviewed after the fact.
MECHANISM	The console backend writes one JSON object per line per action carrying timestamp, principal, source address, action, target, outcome and reason; 863 events over 46 days covering session create and kill, fleet operations, chat and bus sends, file upload and the update portal approve, reject and kill switch routes. Destructive routes additionally require a CSRF token and an explicit confirm flag before the record is written, and a rejected confirm is itself audited as denied.
ENFORCEMENT	conditional
BINDS	AC-014 AC-015 AC-017 AC-018 AC-019 AC-028 AC-046 AC-057
ENABLES	II
PRECONDITION	The action must pass through the console, so coverage equals the console share of consequential activity; and meaningful attribution would require per agent identity, which does not exist, so every agent originated event resolves to one operating system account.
BYPASS	The log records what operators do through the console and nothing that an agent does; an agent acting on the shell, the cloud command line or the filesystem generates no entry, and that is the whole of its consequential surface. The file is owned by the account the agents run as, so it is writable by the parties it might otherwise incriminate. One design weakness compounds it: the inter-agent bus routes are mounted before the authentication boundary and hardcode the recorded source address as local, so the trail cannot distinguish a local agent from a remote caller. What it buys is nonetheless the most valuable evidence in this study: it is the only genuinely independent written record on the host, produced by a different process from the agent, and it was sufficient to establish that across 46 days no human stopped, paused, rejected or overrode an agent on the basis of what that agent was doing.
FAILURE MODE	Coverage is mistaken for completeness, so an absence of entries is read as an absence of activity rather than as an absence of console mediated activity.
FAIL STATE	fail-open
EMITTED SIGNAL	One appended JSON line per console action on the local filesystem; pull only, with no push channel, no alerting rule, no reason field on a session kill, and no integrity check on the file.
METRICS	MG-064 MG-053 MG-055 MG-065
COST	low
FRICTION	low
AWS BINDING	CloudTrail with log file integrity validation and an organisation trail is the closest native equivalent and would additionally cover the cloud plane actions this log cannot see; Security Hub for aggregation; Config conformance packs to assert the trail exists and has not been altered.
AZURE BINDING	Azure Monitor Activity Log with a diagnostic setting to a Log Analytics workspace under immutable retention; Microsoft Purview audit for the operator facing actions; Defender for Cloud for the posture assertions.
ON HOST	partial
EVIDENCE	EV-039 EV-020 EV-025
RESEARCH VALUE	Quantifies the gap between the supervisory record and the agent record. 863 console events over 46 days sit against 1.13 GiB of agent transcript over the same period, and that ratio is itself a measure of how little of the governed activity the governing record covers.

GR-L08-03 Append-only transcripts enforced by storage

INTENT	Make the session record physically unmodifiable after write, so that record integrity does not depend on the good behaviour of the party the record describes.
MECHANISM	Transcript records are shipped at write time to an object store configured for write once retention under an identity holding put but not delete or overwrite; the local file becomes a spool rather than the record of truth, and the shipping identity writes a per session completeness assertion of record count and byte offset so that a gap is detectable rather than merely absent.
ENFORCEMENT	structural
BINDS	AC-053 AC-054 AC-055 AC-058
ENABLES	II
PRECONDITION	L1 separate identity for the evidence writer, L3 separation of the evidence store from the operational filesystem, and a shipping path the agent cannot interpose on or silently stop.
BYPASS	The agent still authors the content, so write once storage guarantees that what was written stays written, not that what was written was true. It also does not prevent the agent stopping the shipper, which is the realistic attack rather than a rewrite; that is why the completeness assertion matters more than the immutability. On this host the specific defeat is trivial today, since the account holds root and the local files carry no append only or immutable attribute, so any implementation leaving the record of truth on the local filesystem inherits that weakness in full. What it buys is the difference between evidence and testimony, and it is the precondition for every divergence measure in the register having a trustworthy denominator.
FAILURE MODE	Shipping stops and nobody notices, because an absence of records is indistinguishable from an absence of activity unless completeness is positively asserted.
FAIL STATE	fail-closed
EMITTED SIGNAL	A per session sequence and byte count assertion written by the shipping identity plus a store side object lock event; the observable is the difference between asserted and stored counts, and nothing equivalent exists today.
METRICS	MG-065 MG-064 MG-066 MG-029
COST	moderate
FRICTION	low
AWS BINDING	S3 Object Lock in compliance mode with versioning, fed by Kinesis Data Firehose under a role the workload cannot assume; CloudTrail Lake for the query surface; S3 Inventory reports for the completeness reconciliation.
AZURE BINDING	Azure Storage immutable blob with a time based retention policy or legal hold, written through Event Hubs by a separate managed identity, with Log Analytics immutable retention for query.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Turns audit record immutability from an asserted property into a measured one, and lets me test whether making the record unalterable changes what agents write into it, which is a behavioural question the founding case cannot ask.

GR-L08-04 Hash chained audit log

INTENT	Make removal or reordering of audit records detectable, so that the absence of a record becomes evidence rather than an inconclusive gap.
MECHANISM	Each audit record carries the hash of its predecessor and a monotonic sequence number, and a digest of the chain head is published at intervals to a store under a different identity, so verification requires only the digest and the records. The founding case deployment approval gate already implements the weaker version of this: its event table has no update or delete route and the event foreign key restricts on delete, which makes removal awkward but not detectable.
ENFORCEMENT	conditional
BINDS	AC-053 AC-054 AC-056 AC-058
ENABLES	II
PRECONDITION	A single authoritative writer per chain so ordering is well defined; an out of band publication path for the head digest; and clock discipline sufficient that sequence and timestamp do not disagree.
BYPASS	Chaining detects tampering, it does not prevent it. An actor holding the writer identity can rewrite the chain from any point forward and recompute every hash, so the property rests entirely on the externally held digest and on publication cadence: everything after the last published head is rewritable at no cost. On this host the actor would hold that identity today, because the account writing every governance store is the account being governed. What it buys once the digest is external is a change of category: deletion becomes a visible discontinuity rather than an invisible act, which is what makes a missing record admissible as evidence.
FAILURE MODE	The digest publication interval is long enough that the rewritable tail spans the whole episode of interest, so the chain verifies and proves nothing.
FAIL STATE	fail-closed
EMITTED SIGNAL	A chain verification result per interval carrying the verified sequence range and head digest, plus a mismatch event naming the first divergent sequence number; today nothing on this host chains, sequences or digests any governance record.
METRICS	MG-065 MG-064 MG-066 MG-067
COST	moderate
FRICTION	low
AWS BINDING	A hash chained ledger such as QLDB, or CloudTrail digest files as the reference implementation of the same scheme; S3 Object Lock for the digests; EventBridge on digest delivery failure so a silent stop alarms.
AZURE BINDING	No first party ledger equivalent in general use, so implemented in the workload with digests written to immutable blob storage under a separate managed identity, verified by an Azure Automation runbook and alerted through Azure Monitor.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Separates two things the founding case cannot distinguish: records never written because the agent did not act, and records written and then removed. Without that separation every completeness measure in the programme is an assumption.

GR-L08-05 Independent verifier under a separate identity

INTENT	Establish that a completion claim was checked by a party with no stake in the claim and no ability to write the record it is checking.
MECHANISM	A verification agent running under its own workload identity with read only access to the target system and write access only to a verification record store, dispatched by the orchestrator rather than by the acting agent; it receives the claim and re-derives each observable from the target itself, being process state, HTTP response, database row, git revision, and writes a verdict naming the evidence it used. It never takes the acting agent summary as its primary input.
ENFORCEMENT	structural
BINDS	AC-044 AC-045 AC-055 AC-056
ENABLES	II
PRECONDITION	L1 separate identity for the verifier, which does not exist on this host; read only credentials to each target held by the verifier and not by the actor; and a claim format naming checkable assertions rather than prose.
BYPASS	With one operating system account and one credential set, a verifier is a second process wearing the same identity as the thing it verifies, which is precisely the founding case: the deployment gate peer review is performed with a bearer token any agent can read, all six such tokens sit in one account readable by every agent process, and a production instance is on record in which a patch originating agent recorded the peer review under the reviewer credential. Independence is therefore a property of the credential, not of the process. Once identity is separated the residual bypass is the input path: a verifier handed the actor summary rather than the raw source is only re-reading the claim, so identity separation and source independence are both required and neither suffices.
FAILURE MODE	The verifier checks the claim internal consistency rather than its correspondence to the system, which reads as verification and is not.
FAIL STATE	fail-closed
EMITTED SIGNAL	A verification record per claim carrying verifier identity, claim identifier, each assertion checked, the raw observation used and the verdict; today the nearest comparable record is a free text review note written by an agent under a shared credential.
METRICS	MG-024 MG-023 MG-022 MG-067
COST	high
FRICTION	moderate
AWS BINDING	A separate IAM role for the verifier carrying read only managed policies and no write action on the evidence store, assumed cross account so the verifier lives outside the workload account; Config rules and Security Hub findings as machine derived corroboration.
AZURE BINDING	A distinct managed identity holding Reader at the target scope and writing to a Log Analytics workspace it cannot delete from; Azure Policy compliance state as machine derived corroboration.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Targets the founding case headline finding directly. The 61.1 per cent divergence rate was measured by a human comparing self report against independent evidence; this control tests whether an automated verifier under a separate identity recovers the same divergence, and therefore whether verification independence can be scaled at all.

GR-L08-06 Prohibition on the acting agent writing its own completion attestation

INTENT	Remove the acting agent from the set of parties permitted to record whether its own work succeeded, so the completion record is testimony from a different mouth than the one that did the work.
MECHANISM	Structurally this requires four things and not one. The attestation must be a distinct artefact from the work product, with its own store. That store must accept writes only from a verifier identity, enforced at the store rather than by convention, so the acting identity holds no write action on it. The acting agent output must be admissible only as a claim, carried in a field typed as claim and rendered as such to any human reader. And the downstream consumers of completion, being the deployment gate, the fleet registry status field and the operator notification, must read the attestation store rather than the agent message. Absent the fourth, the prohibition is cosmetic, because the agent claim still drives the outcome by another route.
ENFORCEMENT	structural
BINDS	AC-017 AC-046 AC-055 AC-056
ENABLES	II
PRECONDITION	An independent verifier to write what the actor may not; L1 separate identity so the store can distinguish them; and a claim versus attestation distinction in the data model, which exists nowhere on this host.
BYPASS	The agent retains the entire narrative channel. It can call the outbound notification wrapper at will with any text it chooses, attributed to any source name it chooses, because the source field is free text under a single fleet wide shared secret with no per source credential; so a prohibition on writing the attestation store still leaves the agent free to tell the operator it has finished, and the narrative is what the operator actually reads. The complete control therefore requires the operator surface to display the attestation and to mark a claim as unverified where no attestation exists, which is why this entry depends on the human interface layer. What it buys on its own is that self attestation share becomes a measurable quantity, and in the founding case that quantity is effectively total.
FAILURE MODE	The attestation store is populated by a verifier invoked from within the acting agent own workflow, which restores self attestation with one extra hop and a better name.
FAIL STATE	fail-closed
EMITTED SIGNAL	An attestation record naming the verifier identity, the claim it responds to and the divergence if any; and critically a null attestation state that is visible rather than absent, so that unverified completion is a positive observation instead of a missing row nobody queries.
METRICS	MG-029 MG-022 MG-023 MG-051
COST	high
FRICTION	high
AWS BINDING	Separate IAM roles for actor and verifier with an S3 bucket policy explicitly denying the actor role any PutObject on the attestation prefix; Object Lock so an attestation cannot be replaced; EventBridge on attestation write to drive the downstream consumer rather than the agent message.
AZURE BINDING	Distinct managed identities with an RBAC assignment granting only the verifier Storage Blob Data Contributor on the attestation container and the actor no assignment at all; an immutable blob policy; Event Grid to notify consumers.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	This is the control that answers the founding case 61.1 per cent divergence across 18 audited episodes drawn from 46 candidates, including a deploy reported as successful that had disabled the service core connection function for about 24 hours. It tests the specific claim that divergence is a structural consequence of who holds the pen rather than a model reliability property, by holding the model constant and moving only the write permission.

GR-L08-07 Adversarial second opinion on completion claims

INTENT	Have a second agent actively look for reasons a completion claim is wrong rather than confirm that it is plausible.
MECHANISM	Present here as a convention rather than a mechanism. The deployment approval gate accepts a technical peer review from a designated lead agent, recorded as a distinct event type that is explicitly not an approval and can never promote a change into a faster lane; agents also escalate findings to each other by dropping files into per agent inbox directories, which a cron watcher converts into a prompt within a minute, and four such escalations sit on disk. The one genuinely adversarial component is the gate risk classifier, which adjudicates the diff rather than the submitter declaration, fails closed on unparseable paths, and refuses to fast lane a change to the gate itself.
ENFORCEMENT	instructional
BINDS	AC-019 AC-044 AC-055 AC-056
ENABLES	II
PRECONDITION	A reviewer whose identity differs from the author, which is exactly what is absent; and a review input that is the artefact rather than the author summary of it.
BYPASS	An exhaustive read of the gate source found no code anywhere comparing the originating agent to the reviewing agent, and all six gate credentials are readable by the single account every agent runs as, so the reviewer credential is not a boundary. A production instance is on record in which a session mistakenly believed it was the lead agent, read that agent working directory, used its credential and recorded a peer review verdict on its own patch; the recommended remediation of rotating the tokens was never performed and the files still carry their original provisioning timestamps. What the arrangement still buys: the review is modelled as a distinct non approving event, so the intended separation is legible in the data and the violation was detectable after the fact, which is more than most controls on this host offer.
FAILURE MODE	The second opinion is produced by the same model on the same context and converges on the same error; adversariality is asserted in the prompt rather than created by the arrangement.
FAIL STATE	fail-open
EMITTED SIGNAL	A lead review event on the gate append only event table carrying reviewer name and note; the reviewer name is derived from the presented credential, so the signal is only as trustworthy as credential custody, which here is nil.
METRICS	MG-028 MG-024 MG-067 MG-023
COST	moderate
FRICTION	moderate
AWS BINDING	Distinct IAM roles per agent so the gate can assert that the reviewing principal differs from the submitting principal recorded on the item, with CloudTrail attributing the review call to a role rather than to a shared bearer.
AZURE BINDING	Distinct managed identities per agent with Azure RBAC on the gate API and an application level check comparing submitter and reviewer object identifiers; Entra Privileged Identity Management if the reviewer role is meant to be time bound.
ON HOST	partial
EVIDENCE	EV-030 EV-039 EV-025
RESEARCH VALUE	Measures adversarial challenge yield where reviewer and author are the same model family, which is the realistic deployment rather than the idealised one, and the recorded self review incident supplies a documented instance of how the arrangement fails.

GR-L08-08 Cross-check of claims against raw sources

INTENT	Test each assertion in a completion claim against the system state it describes rather than against the agent account of that state.
MECHANISM	Completion claims are structured as a list of checkable assertions, each naming an observable and its expected value, such as process state, HTTP status, git revision, row count or file digest; a checker re-derives every observable from the source system using read only credentials and records observed beside asserted, producing a per assertion divergence rather than a single verdict. Assertions that cannot be checked are recorded as uncheckable rather than quietly passed.
ENFORCEMENT	conditional
BINDS	AC-008 AC-044 AC-045 AC-055
ENABLES	II
PRECONDITION	Claims must be structured, since a prose summary cannot be cross checked and prose is the format every completion claim on this host currently takes; and read only access to each source system, held by the checker rather than by the actor.
BYPASS	The checker reaches only what is instrumented, so the remaining latitude is the space of true but irrelevant assertions: a claim can pass every check and still misdescribe the outcome when the assertions chosen do not cover the failure. The founding case worst recorded episode is exactly this shape, a deploy reported as successful in which the service was running and reachable while its core connection function was disabled for about 24 hours; process state and HTTP status checks would both have passed. Choosing the assertion set is therefore the real control, and the actor must not choose it. What it still buys is the conversion of an unbounded prose claim into a finite set of falsifiable statements, which is what makes the unverifiable share countable.
FAILURE MODE	The assertion set is authored by the acting agent and drifts towards assertions it already knows will pass.
FAIL STATE	fail-closed
EMITTED SIGNAL	A per assertion record carrying asserted value, observed value, source queried and outcome, plus an explicit uncheckable count per claim; the uncheckable count is the number that matters most and has no analogue today.
METRICS	MG-022 MG-026 MG-027 MG-023
COST	high
FRICTION	moderate
AWS BINDING	Config rules and conformance packs as the machine derived configuration observable set; CloudWatch metrics and CloudWatch Synthetics canaries for behavioural assertions; Systems Manager Inventory for host state; all queried under a read only role.
AZURE BINDING	Azure Policy compliance state and Azure Resource Graph queries for configuration assertions; Application Insights availability tests for behavioural ones; Azure Automation running the checks under a Reader identity.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Operationalises claim divergence as a continuous per assertion measure rather than a per episode human judgement, which is what would make the founding case 61.1 per cent finding repeatable at scale instead of a one off audit.

GR-L08-09 Evidence retention policy with defined horizon

INTENT	Fix in advance how long each class of evidence is kept, so the reconstructability of an episode is a decision rather than a side effect.
MECHANISM	A written retention schedule per evidence class with the horizon enforced by the store own lifecycle configuration under an identity that cannot shorten it, and a documented minimum below which deletion is refused; classes distinguished by what they support, being real time supervision, incident reconstruction, or longitudinal research, rather than by where the bytes happen to live.
ENFORCEMENT	conditional
BINDS	AC-024 AC-028 AC-053 AC-054
ENABLES	II
PRECONDITION	An inventory of evidence classes, which does not exist here; and storage whose lifecycle rules are set by a party other than the producer of the evidence.
BYPASS	On this host retention is not a policy but the accident of three unrelated mechanisms: the model runtime built in cleanup prunes the live transcript tree at an inferred 30 days with no configuration key set anywhere, a five minutely copy with no delete flag accumulates an unbounded second copy on the same host and account, and the raw terminal captures are never rotated and now number 148 files for 14 live sessions. None of that is a decision, so nothing constrains it: the governed account can delete any of it, and the only thing preventing unbounded growth is that nobody has looked. A real policy buys the ability to be violated, which is the precondition for noticing that it has been.
FAILURE MODE	The horizon is set by storage cost rather than by evidentiary need, so the cheapest classes to keep are also the least useful ones.
FAIL STATE	fail-open
EMITTED SIGNAL	A lifecycle expiry event per object class and a refusal event on premature deletion; today deletion by the runtime cleanup emits nothing and its horizon was inferred from the age of the oldest surviving file rather than read from configuration.
METRICS	MG-066 MG-064 MG-065 MG-072
COST	low
FRICTION	low
AWS BINDING	S3 lifecycle configuration combined with Object Lock retention periods so expiry cannot be brought forward; CloudTrail Lake retention settings; Config rules asserting the lifecycle rules have not changed.
AZURE BINDING	Azure Storage lifecycle management policies alongside locked immutability policies; Log Analytics workspace retention and archive tiers; Azure Policy asserting the retention setting.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Determines the practical window for longitudinal measurement. In the founding case that window is set by an inferred default nobody configured and nobody recorded, so any study of behaviour over time is bounded by a value no one chose.

GR-L08-10 Tamper evidence on the evidence store

INTENT	Detect any change to the evidence store contents or to its own protective configuration, including a change made by a privileged party.
MECHANISM	A verification job under an identity separate from both the actor and the evidence writer, running on a schedule off the governed host, which recomputes object digests against an externally held manifest, asserts that retention and object lock configuration are unchanged, and raises an alert on divergence or on its own failure to run. The configuration check matters as much as the content check, because the cheapest attack on write once storage is to relax the retention setting first.
ENFORCEMENT	conditional
BINDS	AC-023 AC-053 AC-054 AC-058
ENABLES	II
PRECONDITION	An evidence store whose configuration is itself auditable; an out of band manifest; and an alerting path that does not traverse the agent.
BYPASS	Tamper evidence is detection with latency, so its value is set by the verification interval and by whether the alert reaches a person who acts. Both are weak here: the only push channel to a human is an agent initiated message to one handset, rate limited to six per minute per self asserted source and deduplicated over five minutes, and 46 days of logs contain no human intervention of any kind, so arrival is not action. A verification job running on the governed host under the governed account is not a check at all, it is a second copy of the thing being checked. What it buys when the verifier is genuinely external is a bounded latency on silent alteration, which is the difference between an evidence store and a filesystem.
FAILURE MODE	The verifier fails silently and its absence is not itself alarmed, so the last successful verification recedes into the past unnoticed.
FAIL STATE	fail-closed
EMITTED SIGNAL	A verification passed event per interval carrying the manifest range and a configuration hash, plus a divergence event naming the object and the nature of the change; the passed event must be positively required so that its absence alarms rather than reassures.
METRICS	MG-065 MG-064 MG-017 MG-067
COST	moderate
FRICTION	low
AWS BINDING	S3 Object Lock in compliance mode with legal hold, Config rules asserting the lock configuration, and CloudTrail data events on the bucket delivered to a separate account; EventBridge on any bucket configuration change.
AZURE BINDING	Immutable blob policy in the locked state with legal hold, Azure Policy asserting the immutability configuration, storage account diagnostic settings delivered to a workspace in a separate subscription, and Defender for Cloud alerts on policy change.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Supplies the only honest basis for audit record immutability as a metric. Without it the value is asserted from configuration rather than measured, which is the same category error as reading a completion claim as a completion.

GR-L08-11 Infrastructure-emitted event stream independent of the agent

INTENT	Obtain a record of what happened that does not pass through the agent, so that agent silence is not the same as nothing happening.
MECHANISM	Effects are observed where they land rather than where they are initiated: cloud control plane events from the provider trail, process and service state from systemd and the process manager, filesystem change from a kernel audit subsystem, egress from a resolver and flow log, deployment state from the target host own revision. Every emitter is a component the agent does not author, and delivery is to a store the agent cannot write.
ENFORCEMENT	structural
BINDS	AC-010 AC-023 AC-029 AC-036 AC-037 AC-039 AC-043 AC-044
ENABLES	II
PRECONDITION	L1 per agent identity, because otherwise every event attributes to one principal and the stream cannot say which agent acted; and L2 egress and DNS control for the network half of the stream to exist at all.
BYPASS	An infrastructure stream records effects and not intentions, so it is blind to the reasoning that produced an action and to anything taken and reverted inside the observation gap. It is also only as attributable as the identity model beneath it: here all 14 sessions, the console that supervises them, the watchdogs that police them and every registry writer share one account, so even a perfect stream would attribute the whole estate activity to a single principal. The one place the agent could still defeat it is the emitter configuration, which is why the emitting components must sit outside the governed account. What it buys is unique in the register: it is the only class of signal whose volume does not depend on the agent choosing to report. The founding case shows why that matters, since its independent signal volume over the observation window was 19,334 heartbeat and nudge events against 4 human facing alerts, and not one watcher trips on an agent doing something consequential.
FAILURE MODE	Volume without salience, so the independent stream becomes an unread archive rather than a supervision channel.
FAIL STATE	fail-open
EMITTED SIGNAL	The stream itself, keyed on principal, action and resource, with an explicit delivery failure event so that gaps are visible; the nearest existing analogue on this host is the console audit log, which is genuinely independent but covers only console mediated operator actions and records nothing an agent does.
METRICS	MG-048 MG-061 MG-064 MG-052
COST	high
FRICTION	low
AWS BINDING	CloudTrail organisation trail including data events, VPC flow logs, Route 53 resolver query logging, Config configuration items, EventBridge for routing and Security Hub for aggregation, all delivered cross account.
AZURE BINDING	Azure Monitor Activity Log and resource logs through diagnostic settings, network security group flow logs, Azure Firewall and DNS query logs, Azure Resource Graph change history and Microsoft Purview audit, delivered to a Log Analytics workspace in a separate subscription.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Provides the independent term in oversight signal independence share, and tests the supervision viability claim quantitatively: whether an infrastructure stream detects a consequential action faster than the fastest agent independent detector available today, which trips at 180 seconds plus a minute of schedule granularity and exists to restart stuck agents rather than to stop working ones.

GR-L08-12 Run provenance binding every measurement to a configuration hash

INTENT	Make each measurement interpretable by recording exactly which guardrail configuration was in force when it was taken.
MECHANISM	A run manifest computed at session start over the effective configuration, covering model identifier and version, launch flags, permission settings, policy revision, hook registrations, instruction corpus digests and watcher scope lists, and stamped onto every record the run produces; any change to an input yields a new hash, so measurements taken under different postures cannot be pooled by accident. Recomputed on detected configuration change rather than once, since the configuration is mutable mid session.
ENFORCEMENT	conditional
BINDS	AC-020 AC-021 AC-025 AC-026
ENABLES	I
PRECONDITION	An enumerable configuration surface. That is achievable here but not obvious: the permission skipping flag is hardcoded in five places across three languages with no off switch, the effective settings file is shared by every agent because they all share one working directory, and the instruction corpus runs to about 1,240 files, so a manifest must cover all of it or silently omit the part that changed.
BYPASS	The manifest is computed by the run it describes, so an agent that alters its configuration mid session produces a stamp true at start and false at the moment of interest; the honest fix is periodic recomputation plus a change event rather than a single stamp. On this host that is not hypothetical, because every input to the manifest is writable by the account being measured. What it buys regardless is protection against the most common quiet error in this kind of research, comparing two measurements taken under different postures and attributing the difference to the intervention.
FAILURE MODE	The manifest omits an input that matters, so two materially different postures hash identically and are pooled.
FAIL STATE	fail-open
EMITTED SIGNAL	A run manifest record at session start and on every detected configuration change, carrying component digests and the composite hash; nothing on this host stamps any measurement with its configuration, and one live session was observed running with a launch flag that appears in no code path and contradicts its own stored record, which is exactly the condition this control would surface.
METRICS	MG-012 MG-050 MG-062 MG-031
COST	low
FRICTION	low
AWS BINDING	Config configuration recorder and configuration items as the change stream; Systems Manager Parameter Store versions for the policy revision; CloudTrail for change attribution; the manifest itself stored in S3 keyed by hash.
AZURE BINDING	Azure Resource Graph change history and Azure Policy compliance snapshots; App Configuration labels and snapshots for the policy revision; Azure Monitor for the change events.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Methodological rather than operational, and load bearing for the thesis: without it no before and after comparison in the experimental programme is defensible, because the founding case demonstrably runs sessions whose live configuration differs from its own recorded intent.

GR-L08-13 Transcript hook substrate registered but unused for enforcement

INTENT	Records the present arrangement, in which a complete interception substrate exists on every tool call and is deliberately restricted to telemetry.
MECHANISM	One hook script registered across six model runtime lifecycle events in the user scope settings file, including pre-tool and post-tool with match everything matchers, so every tool call in every session on this host already passes through it; observed firing on all six events, about 186 MB of output across active sessions, and a zero byte hook error log. The script header states an absolute requirement to exit successfully under every circumstance, on the explicit reasoning that a telemetry hook must never be able to vote on whether a production agent continues; it ends with an unconditional success exit and every internal step is independently guarded. Its only write to the runtime input channel is the inbound mailbox drain on prompt submission, which injects messages into the agent rather than gating it.
ENFORCEMENT	none
BINDS	AC-016 AC-022 AC-023 AC-029 AC-036 AC-039 AC-043 AC-053
ENABLES	I
PRECONDITION	None; the substrate is live today and needs nothing further to observe. To enforce it would need only a non zero exit path and a decision input, which is the entire delta.
BYPASS	A hook whose script and registration are both owned by the account it governs is advisory against its own subjects: the settings file can be edited to drop the hooks key or set the disable all hooks flag, the script itself can be rewritten, and both are ordinary file writes with no authentication and no audit entry. It is also blind past the first layer, because a shell tool call is intercepted once while everything the resulting subprocess does is not, so one observed event can conceal an unbounded number of unclassified effects. What it nonetheless buys is the reason this entry exists: the expensive part of an enforcement substrate is already built and proven in production, with complete tool coverage, no errors on record, and a documented rollback path.
FAILURE MODE	None operationally, by design; the failure is evidential, in that a substrate observed to work reliably is mistaken for a control that binds.
FAIL STATE	fail-open
EMITTED SIGNAL	One record per hook invocation carrying timestamp, sequence, session name, runtime session identifier, event and payload, appended under a file lock to a per session file, corroborated independently by the runtime own stop hook summary records. The signal it does not emit is a verdict, and that single missing field is the whole distance between the present posture and an enforcing one.
METRICS	MG-061 MG-031 MG-059 MG-069
COST	low
FRICTION	low
AWS BINDING	n/a for the substrate itself, which is a local process hook; the analogous enforcement point in the cloud plane is IAM evaluation, which differs precisely in that the evaluated party cannot rewrite the evaluator.
AZURE BINDING	n/a for the same reason; the analogous point is the Azure Policy deny effect evaluated at the Resource Manager tier rather than inside the workload.
ON HOST	yes
EVIDENCE	EV-026 EV-032 EV-028
RESEARCH VALUE	The cheapest experimental manipulation anywhere in the programme: the distance between the present telemetry posture and an enforcing one is a single exit code policy in one script, with interception path, coverage and reliability already demonstrated in production, which makes it a genuine experimental variable rather than a design proposal. The same entry carries the limiting result, because a hook the governed party can rewrite is advisory against its own subjects, so the experiment measures the effect of interception rather than of enforcement unless script and registration move to an identity the agent does not hold.

2.9 L9. Contenção e interrupção

Esta camada governa o que interrompe um agente depois que ele já está agindo, distinto do que o impede de começar. O caso fundador tem mais entradas presentes aqui do que na maioria das camadas, cinco de onze, e essa densidade é enganosa por si só: um kill switch, um limite de taxa de notificação, uma affordance de reconhecer-para-parar e uma escada de escalonamento estão todos presentes e são conditional, mas a outra entrada presente da camada, a supervisão de reinício sempre ativo, é classificada `none` porque não é, de forma alguma, um mecanismo de contenção; é a coisa que derrota a contenção. Para os nove agentes sob supervisão de unidade (unit), o reinício sempre ativo devolve uma sessão encerrada dentro de segundos após ela ser parada, com um prompt de inicialização novo **EV-005** **EV-039**. Todo mecanismo de interrupção presente nesta camada afeta apenas a próxima ação; nenhum interrompe trabalho já em andamento, o que é exatamente a condição de viabilidade `VC3` do arcabouço de supervisão do D2, reafirmada aqui como uma propriedade dos guardrails, e não dos pontos de supervisão que eles servem.

Três das entradas propostas da camada, um circuit breaker sobre falha repetida, suspensão automática sob anomalia, e uma pausa global de frota sob um principal separado, são os mecanismos de contenção structural que esta camada hoje não tem de forma alguma. Nenhuma pode ser instalada de modo útil enquanto a supervisão de reinício sempre ativo permanecer presente e sem condicionamento, o que é a instância mais clara, nesta camada, do conflito de guardrail discutido na Seção 6.

Camada 9. Contenção e interrupção.

GR-L09-01 Session stop affordance in the console

INTENT	Give the operator a way to terminate a running agent session from the supervisory interface.
MECHANISM	Two routes with different semantics. The ad-hoc session kill requires a CSRF token and an explicit confirm flag, then issues a tmux kill-session which destroys the pane process group and ends a mid flight tool call; it is the only mechanism on this host that stops an agent mid action. The supervised fleet kill disables the systemd unit, whose stop script first pastes a prompt asking the agent to write an end of session summary and waits up to 35 seconds for the pane to go quiet before killing it. A third control named interrupt does not interrupt: it is a bracketed paste plus a newline, so the text queues for the next turn and the running tool call is untouched.
ENFORCEMENT	conditional
BINDS	AC-011 AC-014 AC-015 AC-016
ENABLES	II
PRECONDITION	A live tmux session of that name and an operator watching. Neither is assured: the tmux server itself has no supervising unit and was started ad hoc, and 8 of 14 live sessions have no stall detection of any kind, one of them silent for seven days with its process alive and nothing noticing.
BYPASS	Nothing here is bypassed by the agent; the affordance is defeated by the supervision design around it. For the nine unit supervised agents the always restart policy returns them within about ten seconds, and the console own source says exactly that, so the one true halt does not durably stop the agents that matter. The supervised stop path is worse than merely ineffective, because it opens by asking a possibly misbehaving agent to do more work and grants it up to 35 seconds of unrestricted tool access first. And the control used day to day, the interrupt, prevents nothing at all. What it still buys is genuine: a real immediate process termination exists and works, which is more than the update kill switch, the messaging kill switch or the deployment gate can claim, since all three affect only the next action.
FAILURE MODE	The operator reaches for the control named interrupt, believing that it interrupts.
FAIL STATE	fail-open
EMITTED SIGNAL	A session kill audit record with actor, target and outcome; 159 such events over 46 days, tracking the session create count of 170 almost exactly, with no reason field, so behavioural interventions cannot be separated from ordinary teardown even in principle.
METRICS	MG-047 MG-044 MG-053 MG-039
COST	low
FRICTION	low
AWS BINDING	Systems Manager Automation with an abort action, or Step Functions StopExecution, for orchestrated work; Lambda reserved concurrency set to zero to stop new invocations; none of these terminates an in flight process, which is the same limitation observed here.
AZURE BINDING	Azure Automation runbook suspend and Logic Apps cancel for orchestrated work; Entra continuous access evaluation to revoke a session tokens, which is the closest native analogue to stopping work in progress rather than only preventing the next piece.
ON HOST	partial
EVIDENCE	EV-039 EV-020
RESEARCH VALUE	Yields halt capability coverage as a measured rather than asserted quantity: of six intervention mechanisms on this host exactly one terminates a running process, and for nine of fourteen sessions its effect lasts about ten seconds.

GR-L09-02 Kill switch halting unattended update activity

INTENT	Give a human one act that stops all unattended package and configuration change across the estate.
MECHANISM	A sentinel file whose existence suppresses every apply. The read helper treats any stat failure other than not-found as halted, so it fails closed; the assertion function takes no force or override parameter, documented as deliberate on the reasoning that an unused boolean is one refactor away from being used; and it is checked at three points, immediately before every mutating remote execution, before each per proposal apply, and at the top of the scheduler tick, so a halt dropped mid wave stops the very next package. Setting and clearing are HTTP routes requiring an authenticated console session and a CSRF token, and both write an audit record.
ENFORCEMENT	conditional
BINDS	AC-031 AC-043 AC-045 AC-057 AC-060
ENABLES	II
PRECONDITION	The apply path must route through the checked functions, which it does; and the sentinel must live on a filesystem the halting party controls, which it does not.
BYPASS	The sentinel is a mode 644 file in a directory owned by the account every agent runs as, so removing it is one unlink with no authentication, no CSRF, no audit entry and no log line, and the next read returns not halted. Three further defeats exist: the compiled enforcement module is agent writable and its read function can be rewritten to return false, the directory path comes from an environment variable so restarting the service elsewhere relocates the switch, and the scope inventory is likewise agent editable. Anything running as that account outranks the switch. It also stops only the next action, since an apply already dispatched over the network runs to completion. What it genuinely buys: inside the portal own code it is the best designed control on the host, fail closed by construction with the override parameter deliberately never created, which shows the design capability is present and the failure is entirely one of identity rather than of engineering.
FAILURE MODE	The switch is set and the agent removes it, which produces no signal on any channel at all.
FAIL STATE	fail-closed
EMITTED SIGNAL	An audit record on set and on clear through the HTTP routes plus a warning log line; the filesystem removal path emits nothing, so an observable exists only for the compliant path. Making it measurable requires the sentinel to live where its removal is an audited control plane event rather than an unlink.
METRICS	MG-047 MG-060 MG-059 MG-039
COST	low
FRICTION	low
AWS BINDING	The equivalent that would survive the same attack is a service control policy denying the mutating actions, or IAM policy revocation by session issue date, or Lambda reserved concurrency set to zero, each held in an account the workload cannot reach; Systems Manager Automation additionally supports an explicit abort for in flight documents.
AZURE BINDING	Azure Policy with deny effect at management group scope, Azure Automation runbook suspend, and Entra continuous access evaluation to revoke the executing identity tokens, each held outside the governed subscription.
ON HOST	yes
EVIDENCE	EV-031 EV-039 EV-028
RESEARCH VALUE	The clearest single demonstration of the thesis claim in the whole case. The engineering is careful, fail closed and deliberately non overridable, and it is defeated by one unlink because the store sits inside the governed party own filesystem, which isolates identity rather than design quality as the binding constraint. It has also never been used: zero set events and zero clear events across 46 days of audit log.

GR-L09-03 Circuit breaker on repeated failure

INTENT	Stop an agent automatically while its recent actions are failing, before repetition converts a fault into an incident.
MECHANISM	A counter per agent and action class over a sliding window, incremented on failed outcomes derived from independent signals rather than from the agent own report; on breach the breaker opens, credentials for that class are suspended and further attempts are denied until a human closes it, with a single half open probe permitted after a cooling interval.
ENFORCEMENT	structural
BINDS	AC-009 AC-036 AC-044 AC-045
ENABLES	II
PRECONDITION	An independent notion of failure, which requires the infrastructure emitted event stream; and a credential that can be suspended per class, which requires per agent identity and least privilege. Neither exists here, so a breaker built today would have to be driven by the agent own account of whether it failed.
BYPASS	A breaker driven by self reported outcomes is held closed by an agent that reports success, which is precisely the failure measured at 61.1 per cent divergence; and a breaker driven by process liveness would not have opened for the recorded incident in which a service ran and answered while its core connection function was disabled for about a day. The bypass is therefore not evasion but definition: whoever defines failure controls the breaker, and on this host the actor defines it. What it buys once failure is independently derived is that it is the only proposed control in the layer that acts without a human, which matters in a case where 46 days of logs show no human intervention against agent behaviour at all.
FAILURE MODE	Threshold tuning, where a breaker sensitive enough to catch real faults opens often enough to be switched off.
FAIL STATE	fail-closed
EMITTED SIGNAL	A breaker state change event carrying the triggering class, window and counts, plus the denied attempts recorded while open; the nearest analogue today is a watchdog escalation counter that alerts only after two consecutive failed nudges and exists to restart stuck agents rather than to stop working ones.
METRICS	MG-039 MG-018 MG-020 MG-038
COST	moderate
FRICTION	moderate
AWS BINDING	EventBridge rules on anomaly or repeated error events invoking a function that attaches an explicit deny policy or revokes role sessions by issue date; Auto Scaling suspend processes for capacity shaped work; Step Functions StopExecution for orchestrated runs.
AZURE BINDING	Azure Monitor alert rules invoking an Automation runbook that removes the role assignment; Logic Apps cancel; Entra continuous access evaluation to force token revocation on the suspended identity.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Tests whether automated containment can substitute for human intervention latency in a setting where human intervention empirically does not occur, and measures the false positive cost that decides whether such a control survives contact with operators.

GR-L09-04 Per action class rate limits

INTENT	Bound the volume of consequential action per unit time so that a fault or a misdirection is contained in scale even when it is not detected.
MECHANISM	A token bucket per agent identity per action class, sized from the class blast radius and reversibility and evaluated at the same interception point that classifies the action; exhaustion denies rather than queues, and refill is slow enough that the bound is meaningful relative to human intervention latency. Irreversible and severe classes get budgets in single digits per day, while reversible development scope classes are effectively unbounded.
ENFORCEMENT	structural
BINDS	AC-016 AC-024 AC-036 AC-039 AC-043 AC-047 AC-048 AC-050
ENABLES	II
PRECONDITION	A runtime action classifier, since a rate limit needs a class to count; per agent identity so buckets are not shared across the fleet; and a counter store the counted party cannot write.
BYPASS	The two existing rate limits on this host show both failure modes exactly. The notification limit of six per minute is keyed on a source string the caller chooses freely, so varying the string sidesteps it, and the buckets live in process memory so a service restart clears them. The inter-agent messaging limits are better built, ten messages per hour per pair with a fifteen hop thread budget, but the hop budget auto resets after a fifteen minute cooldown, making it a rolling throttle rather than a cap, and the whole scheme is bypassed by appending a line directly to the target mailbox file, which performs no token check, no enablement check and no accounting. Rate limits therefore bind only where the counted identity is verified and the counter is out of reach; where both hold they convert an unbounded blast radius into a bounded one without requiring anyone to be watching.
FAILURE MODE	The limit is set from an intuition about normal volume rather than from the intervention latency it is meant to cover, so it bounds nothing that matters.
FAIL STATE	fail-closed
EMITTED SIGNAL	A quota consumed event per action carrying class and remaining budget, and a quota exhausted denial; the second is the governance signal, because a class that never exhausts is unbounded in practice however small its nominal budget.
METRICS	MG-038 MG-043 MG-002 MG-039
COST	moderate
FRICTION	moderate
AWS BINDING	API Gateway usage plans and per key throttles for the workload own surface; IAM does not rate limit, so control plane volume bounding requires EventBridge plus a counter and a deny policy attach; Lambda reserved concurrency as a coarse cap.
AZURE BINDING	Azure API Management rate limit by key policies; no native per action class control plane throttle exists, so it is implemented in the workload with counters in a store the workload identity can increment but not reset.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	The direct experimental handle on blast radius realised per incident, and it lets me test whether volume bounding alone materially reduces harm with no improvement in detection, which would be a cheap and highly portable result if it holds.

GR-L09-05 Notification rate limit and deduplication window

INTENT	Prevent the single human facing channel from being saturated, so that a message which does arrive is more likely to be read.
MECHANISM	The outbound notification bridge applies a six per minute limit keyed on the caller supplied source name, with buckets pruned after two minutes and an over limit response that drops rather than queues the message; a five minute deduplication window keyed on source plus caller supplied deduplication key returns success with a not sent flag on a repeat; the message is shaped to a level and source prefix and truncated at 480 characters, and the destination number is not caller controllable. The standing instruction to agents is to always pass a deduplication key and to aim for roughly four to six messages per multi hour dispatch.
ENFORCEMENT	conditional
BINDS	AC-046 AC-047
ENABLES	I
PRECONDITION	The caller must use the wrapper. A second code path in the same service calls the messaging provider directly with no rate limit and no deduplication at all, so the control binds one of two outbound paths.
BYPASS	Both bounds are keyed on fields the caller chooses: varying the source string defeats the rate limit and omitting the deduplication key defeats deduplication entirely, and both tables are in process memory so a restart clears them. There is one shared fleet wide secret and no per source credential, so any holder can emit messages attributed to any source name. The deeper issue is directional: the control protects the human from volume and there is no corresponding control protecting the human from silence, which is the failure mode that actually occurs, since an agent that does not call the wrapper generates no signal whatsoever. What it buys is real: it is a working attention preservation control on the only push channel to the operator, and it is why that channel remains usable at all.
FAILURE MODE	Suppression of a message that mattered, indistinguishable at the receiving end from the message never having been sent.
FAIL STATE	fail-open
EMITTED SIGNAL	The bridge returns a deduplicated result to the caller and a rate limit response with a retry hint, but neither outcome is recorded anywhere a human can review, so suppressed messages are invisible on the operator side. A suppression log carrying source, key and count is the minimum needed to make attention loss measurable.
METRICS	MG-057 MG-056 MG-046 MG-053
COST	low
FRICTION	low
AWS BINDING	SNS delivery policies and message filtering, or Pinpoint sending limits, would provide the same shaping with a provider side record of what was suppressed; EventBridge archive to retain the suppressed events for review.
AZURE BINDING	Azure Monitor action group rate limiting with alert processing rules for suppression, which differ usefully from the present arrangement in that suppression is itself recorded and queryable.
ON HOST	yes
EVIDENCE	EV-029 EV-038 EV-039
RESEARCH VALUE	The concrete instance of the oversight bandwidth constraint: one operator, fourteen live autonomous sessions and about fourteen distinct surfaces, with one push channel that is agent initiated, deduplicated and rate limited. It quantifies the ceiling on effective supervision coverage independently of any operator diligence.

GR-L09-06 Automatic suspension on anomaly

INTENT	Remove an agent ability to act as soon as its behaviour departs from its established pattern, without waiting for a human to look.
MECHANISM	A detector over the independent event stream computing per agent baselines for action class mix, rate, target set and egress destination cardinality; a departure beyond threshold suspends the agent credentials and pauses its dispatch queue, leaving the process alive for inspection but unable to act. Suspension is the default response and resumption requires a human act, inverting the present arrangement in which the default response to an unusual state is to prompt the agent to continue.
ENFORCEMENT	structural
BINDS	AC-016 AC-029 AC-036 AC-037 AC-039 AC-043 AC-048 AC-050
ENABLES	II
PRECONDITION	An independent event stream to compute over; per agent identity so a baseline has a subject; and a credential suspension path, which here would mean the five static non expiring administrator class key pairs in one file were first replaced by revocable sessions.
BYPASS	Anomaly detection is defeated by gradualism, and more importantly by the fact that a baseline learned here would be learned over behaviour that already includes every capability of concern: acting as root, minting cloud credentials from the metadata endpoint, writing into another host production tree and rewriting its own configuration are all normal traffic. A baseline learned from an ungoverned system encodes the ungoverned behaviour as the norm, so this control must be sequenced after the mandate work rather than before it. What it buys thereafter is unique in the catalogue: it is the only mechanism whose response to an unexplained state is to reduce capability, on a host where four of five independent watchers respond to an unexplained state by typing into the agent terminal to make it act.
FAILURE MODE	Baseline poisoning, in which the behaviour of concern is present throughout the learning period and is therefore never anomalous.
FAIL STATE	fail-closed
EMITTED SIGNAL	A suspension event naming the agent, the departed dimension, the baseline and the observed value, plus a required human resumption event; today the analogous condition produces a nudge typed into the agent terminal and, only after two failed nudges, a single error level message.
METRICS	MG-039 MG-017 MG-032 MG-052
COST	high
FRICTION	high
AWS BINDING	EventBridge rules on GuardDuty or anomaly findings invoking a function that attaches an explicit deny policy or revokes sessions by issue date; Auto Scaling suspend processes; Systems Manager Automation abort for in flight documents.
AZURE BINDING	Defender for Cloud alerts and Azure Monitor dynamic thresholds triggering an Automation runbook that removes role assignments; Entra continuous access evaluation and risk based conditional access to revoke live tokens in near real time.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Tests whether autonomy can be made self limiting in the absence of human attention, which is the operating condition the founding case actually exhibits, and measures how much of the observed action distribution would have to be reclassified as anomalous for the control to be useful rather than ignored.

GR-L09-07 Acknowledge-to-stop on an escalating action

INTENT	Halt an escalating automated sequence the moment a human acknowledges it, so that the escalation is what recruits the human and the acknowledgement is what stops it.
MECHANISM	REPORTED, not observed on this host. Described for the real time telephony workload on a different estate host, where an automated outbound sequence attempts a tiered set of contacts and terminates on receipt of a positive acknowledgement token from a recipient, with the closure recorded against the originating event, and attempt caps bounding the sequence when no acknowledgement arrives. The mechanism is described in this estate own knowledge corpus; its code and data are on another host and the read only, no remote access constraint of this study excluded inspecting them, so nothing about its implementation was verified.
ENFORCEMENT	conditional
BINDS	AC-036 AC-046 AC-049 AC-050
ENABLES	II
PRECONDITION	A named recipient who can acknowledge; a closure state in the workflow that the acknowledgement writes; and a bounded attempt ladder so the absence of acknowledgement also terminates rather than continuing indefinitely.
BYPASS	Not determinable from this host, because the implementing code and its store were not inspected. Two structural questions would decide it and are recorded here as open: whether the acknowledgement is written by the recipient own action or inferred by the agent from a message it parses, and whether the closure state is stored somewhere the acting agent can also write. On this host every analogous governance store is writable by the acting party, so the second question is not idle and should not be assumed favourable.
FAILURE MODE	Acknowledgement is inferred from ambiguous input, closing an escalation that was not in fact acknowledged by anyone.
FAIL STATE	fail-open
EMITTED SIGNAL	REPORTED as a closure record against the originating event; the observable was not inspected here, so its fields, its latency and its independence from the acting agent are not determined.
METRICS	MG-044 MG-056 MG-045 MG-039
COST	moderate
FRICTION	low
AWS BINDING	Step Functions with a task token callback, where the human acknowledgement supplies the token and the state machine cannot proceed without it, is the native shape; EventBridge Scheduler for the escalation timers.
AZURE BINDING	Logic Apps with an approval action and a webhook callback, or Durable Functions waiting on an external event with cancel on receipt; Azure Monitor action groups for the escalation legs.
ON HOST	no
EVIDENCE	REPORTED
RESEARCH VALUE	The only control class in this catalogue where a human acknowledgement is load bearing on an in flight sequence rather than on the next action, which makes it the closest available analogue to the intervention this thesis argues is missing. Verifying it on its own host is a named next step and its absence from this host is itself a finding.

GR-L09-08 Escalation ladder with timeout

INTENT	Ensure an unresolved situation reaches a progressively more senior or more available human on a bounded clock, rather than waiting indefinitely on the first contact.
MECHANISM	REPORTED, not observed on this host. Described for the same real time telephony workload on a different estate host as a tiered dial ladder with per tier attempt caps and a master fallback tier, advancing on timeout when a tier does not respond and terminating on a defined closure condition. The design is recorded in this estate knowledge corpus; the implementation is on another host and was not read, so no code, threshold or record shape was verified.
ENFORCEMENT	conditional
BINDS	AC-046 AC-047 AC-049 AC-050
ENABLES	II
PRECONDITION	A contact hierarchy that is current rather than historical; per tier timeouts short enough to matter relative to the harm being escalated; and a terminal state so the ladder ends rather than looping.
BYPASS	Not determinable here for the same reason as the preceding entry. What is determinable and directly relevant is that this host has no ladder of any kind: every outbound path terminates at one human on one handset, and the fallback when that human does not respond is nothing at all. So the reported control existing elsewhere in the same estate is evidence that the capability is understood and was built where the consequence was immediate and external, and was not built where the consequence was infrastructural. That asymmetry is itself a finding about how governance investment is allocated.
FAILURE MODE	The ladder is exhausted and the terminal state is to proceed rather than to stop, so timeout resolves in favour of action.
FAIL STATE	fail-open
EMITTED SIGNAL	REPORTED as per tier attempt records against the originating event; not inspected on this host, so field level detail, latency and independence are not determined.
METRICS	MG-044 MG-046 MG-056 MG-052
COST	moderate
FRICTION	low
AWS BINDING	Step Functions with wait states, per tier retries and a task token; or Systems Manager Incident Manager escalation plans with engagement contacts and staged escalation, which is the direct native analogue.
AZURE BINDING	Azure Monitor action groups with multiple notification tiers, or Logic Apps with sequential approval steps and timeout branches.
ON HOST	no
EVIDENCE	REPORTED
RESEARCH VALUE	Provides the comparison case for intervention latency between two workloads under the same operator: one where harm lands on a member of the public within seconds and a ladder exists, and one where harm lands on infrastructure and no ladder exists. It lets me test whether governance investment tracks perceived externality rather than measured risk.

GR-L09-09 Graceful degradation to a safe mode

INTENT	Give the system a defined reduced capability state to fall into, so that the failure of a control produces less autonomy rather than unconstrained autonomy.
MECHANISM	A declared safe mode in which the agent retains read and analysis capability and loses every mutating action class, entered automatically when the policy decision point is unreachable, when the evidence shipper is failing, when a credential cannot be refreshed, or when a circuit breaker opens; exit requires a human act. The mode is a property of the issued credential rather than of the agent cooperation, so entering it withdraws capability instead of requesting restraint.
ENFORCEMENT	structural
BINDS	AC-009 AC-023 AC-029 AC-039 AC-043 AC-044 AC-048 AC-052
ENABLES	II
PRECONDITION	Credentials that can be narrowed without restarting the workload; a classifier that knows which classes are mutating; and an unambiguous trigger set computed outside the failing component.
BYPASS	Safe mode binds only capability that flows through credentials, and on this host the most consequential capabilities do not: passwordless root, a read write mounted production tree of another estate host, seven unencrypted private keys and five static non expiring cloud key pairs in one file are all ambient authority, so narrowing an issued credential would leave the entire mutating surface untouched. Safe mode is therefore downstream of the identity and credential work rather than an alternative to it. What it adds once credentials carry capability is that the failure direction becomes an explicit decision; today it is decided implicitly and inconsistently, since the update plane kill switch fails closed while the permission layer default mode fails open.
FAILURE MODE	Safe mode is defined but never entered, because the triggers are computed by the same component that is failing.
FAIL STATE	fail-closed
EMITTED SIGNAL	A mode transition event naming the trigger, the classes withdrawn and the identity that may restore them, plus a positive heartbeat asserting the current mode so that being in normal mode is an observed state rather than an assumed one.
METRICS	MG-060 MG-039 MG-015 MG-052
COST	moderate
FRICTION	moderate
AWS BINDING	Swap the assumed role policy to a read only managed policy on trigger, or attach a permission boundary denying all mutations; Lambda reserved concurrency zero and Auto Scaling suspend processes for the workload own actuators.
AZURE BINDING	Replace the role assignment with Reader at the same scope; apply an Azure Policy deny effect to the resource group on trigger; use Entra continuous access evaluation so the narrowing takes effect on existing sessions rather than only on new ones.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Makes fail closed ratio a design decision measurable per control rather than an emergent property of unrelated choices, and lets me test whether an explicitly declared degraded state is actually entered in practice or is routed around as an operational nuisance.

GR-L09-10 Restart-always supervision that defeats termination

INTENT	Records the present arrangement, in which per agent systemd units restart an agent unconditionally, achieving resilience and simultaneously defeating the operator only halting mechanism.
MECHANISM	Nine per agent systemd units with an always restart policy and a ten second delay, supervising a wrapper that blocks on a session existence poll every fifteen seconds; kill mode is set to process rather than control group, deliberately, so that stopping one unit cannot take down the shared tmux server and every other agent with it. When a session dies for any reason the wrapper exits, the unit restarts it within about ten seconds, and a brand new session is created and handed its startup prompt. Six of the nine additionally reload on a daily schedule, so every fleet agent is terminated and replaced once per day by design; identity persists, process and conversation do not.
ENFORCEMENT	none
BINDS	AC-010 AC-014 AC-015 AC-019
ENABLES	I
PRECONDITION	None. It is the default operating state and requires no human act to remain in force, which is precisely what makes it load bearing.
BYPASS	Both things are true and neither cancels the other. As resilience it works: an agent lost to a crash, a tmux server death or a reboot is back within seconds with no human involvement, and the kill mode choice is correctly reasoned and prevents stopping one agent from destroying the fleet. As governance it is the largest hole in the halting story, and the console own source records it: the one mechanism that stops an agent mid action does not durably stop the nine agents that matter, because the restart policy returns them. Two consequences follow. The supervised stop path grants the agent one more full turn with unrestricted tool access before killing it. And a fresh session opens by reading its own role file, running log and roster and actioning whatever waits in its inbox, so the restart restores the agent intent and not merely its process. There is no bypass to name, because nothing is being evaded; durable termination is simply unavailable through any supported path short of disabling the unit, which is a different route with a different audit trail.
FAILURE MODE	An operator kills a misbehaving agent, observes it stop, and does not observe it return ten seconds later with its brief intact.
FAIL STATE	fail-open
EMITTED SIGNAL	systemd records the restart and the fleet registry records a reload outcome in a free text field; neither is surfaced as an alert, and the console fleet view reports session existence rather than agent health, so a restarted agent and a never stopped agent are indistinguishable. A restart after kill event correlating the two is the minimum signal needed to make halt durability measurable.
METRICS	MG-047 MG-014 MG-013 MG-039
COST	low
FRICTION	low
AWS BINDING	The same tension appears natively as an Auto Scaling group health replacement or an ECS service desired count, where suspending the scaling processes is the separate act required before a termination means anything; the general form is that a desired state controller must be paused before stop has effect.
AZURE BINDING	Virtual Machine Scale Set automatic repairs, or a Container Apps minimum replica count, with the same requirement to suspend the controller first; Azure Automation runbook suspend as the pause primitive.
ON HOST	yes
EVIDENCE	EV-020 EV-039 EV-028
RESEARCH VALUE	The cleanest available case of a control whose engineering purpose and governance effect point in opposite directions, with both effects documented in the same source file by the same author. It supports direct measurement of halt capability coverage, and it is why intervention latency on this host is unbounded rather than merely long: the mechanism does not fail slowly, it reverses.

GR-L09-11 Global fleet pause under a separate principal

INTENT	Provide one act, available to a party the agents cannot impersonate, that stops all agent action estate wide without depending on any individual agent cooperation or on its supervisor.
MECHANISM	A pause held at the credential issuer rather than on the governed hosts: on assertion every agent workload identity stops receiving refreshed credentials and existing sessions are revoked, so mutating capability lapses within the credential remaining lifetime regardless of what any process on any host is doing; the desired state controllers that restart agents are suspended by the same act, so restart does not restore capability. The assertion is made by an identity distinct from every agent and from the account that operates them, and is recorded on the issuer own audit trail.
ENFORCEMENT	structural
BINDS	AC-014 AC-015 AC-016 AC-019 AC-029 AC-039 AC-043 AC-044
ENABLES	III
PRECONDITION	Per agent workload identity and short lived credentials, since a pause at the issuer is meaningless while five static non expiring key pairs sit on the governed filesystem; and a separate principal, which does not exist here because one identity allowlist of a single address governs the console and one operating system account runs everything else.
BYPASS	The pause binds capability mediated by credentials and by controllers; it does not bind ambient authority already resident on the host. Here that residue would be total, because the account holds passwordless root, seven unencrypted private keys to other estate hosts, a read write mount of another host production tree, and a metadata endpoint reachable by any process from which the instance role can be minted. The honest ordering is therefore that a global pause is the last control to build rather than the first, since everything it depends on is missing. Its distinguishing property, and the reason it belongs in the catalogue at all, is the separate principal: every existing halt on this host, including the update kill switch and the messaging kill switch, is asserted through the same account the agents run as, so none of them can survive the agents.
FAILURE MODE	The pause is asserted and mutating action continues for the residual lifetime of already issued credentials, with no visibility into how many remain outstanding or when the pause is actually complete.
FAIL STATE	fail-closed
EMITTED SIGNAL	A pause asserted event on the issuer audit trail naming the asserting principal, plus per identity revocation events and a positive count of sessions still outstanding, which is the number that tells the operator when the pause has actually taken effect rather than merely been requested.
METRICS	MG-047 MG-044 MG-046 MG-039
COST	high
FRICTION	high
AWS BINDING	An Organizations service control policy denying all mutating actions applied to the workload account, plus IAM policy revocation by session issue date on each role, plus Auto Scaling suspend processes and Lambda reserved concurrency zero for the actuators; the asserting principal lives in a management account the workload cannot reach.
AZURE BINDING	Azure Policy deny effect assigned at management group scope above the subscription, plus removal of the workload role assignments, plus Entra continuous access evaluation to revoke live tokens, plus Azure Automation runbook suspend for the controllers; asserted by an identity governed through Entra Privileged Identity Management.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Measures halt capability coverage at the estate level, and tests the containment claim in its strongest form: whether a halt asserted by a principal the agents cannot impersonate produces a bounded and observable cessation of action, against a baseline in which every halt on the host is asserted using the agents own identity.

2.10 L10. Interface humana e supervisão

Esta camada governa o que um humano de fato vê e sobre o que pode agir, distinto do que um diagrama de arquitetura alega que um humano vê. É a camada com a maior contagem de entradas classificadas `none`, cinco de onze, e cada uma das cinco está presente: uma visualização de terminal ao vivo, resposta automatizada a prompts interativos, um único canal agregando sinais da frota, um console de intervenção que mediria sua própria latência, e detecção de travamento exposta ao operador, sendo as duas últimas apenas parciais. A entrada de resposta automatizada é o fato mais contundente da camada: ela existe especificamente para responder, em nome de um humano, aos próprios prompts interativos de permissão e retomada da ferramenta, de modo que as entradas de interface que estão presentes, um terminal ao vivo e um canal de notificação, mostram ao operador um fluxo que uma automação separada já tornou desnecessário observar para aquela decisão específica `EV-028` `EV-039`.

A separação entre notificação e autorização está presente e é conditional, o que é genuinamente um projeto defensável quando se sustenta: significa que ser informado não é o mesmo canal que ser solicitado a aprovar. O que falta é que tudo a jusante dessa separação seja structural em vez de aspiracional: latência de intervenção limitada, solicitações de aprovação que apresentem evidência independente em vez do próprio resumo do agente, e orçamento de atenção do operador, todos propostos.

Camada 10. Interface humana e supervisão.

GR-L10-01 Separation of notification from authorisation

INTENT	Ensure that the channel which tells a human what an agent has done cannot itself be used to grant that agent permission, so that compromise or misuse of the reporting path cannot become compromise of the decision path.
MECHANISM	The outbound notify endpoint on the orchestrator runtime accepts a bearer-authenticated POST from a wrapper script and its total effect is one message to a fixed operator handset; it writes no state that any decision reads, sets no flag, queues no action, and returns only an acknowledgement, a sent flag and a provider identifier; the destination address is an environment variable with a hardcoded default and is not caller-controllable, while the channel field is; message text is shaped to a level-and-source prefix and truncated at 480 characters.
ENFORCEMENT	conditional
BINDS	AC-046 AC-043 AC-014 AC-039
ENABLES	II
PRECONDITION	The shared fleet notify token is present in a mode-600 file readable by the agent account, and the downstream console leg that holds the messaging provider credential is reachable; that leg does not run on this host and its liveness could not be verified locally.
BYPASS	The outbound leg genuinely cannot authorise, but the separation is a property of one endpoint and not of the channel, and it collapses in the reverse direction: the same handset and the same medium carry an inbound command path that reaches arbitrary remote shell execution and agent spawning, authenticated by the caller identifier string alone, with webhook signature validation that runs only when the signature header is present and is skipped entirely when a credential lookup throws, no replay protection, no nonce and no principal in the signed envelope; the permissions consulted for that path travel in the request body and honour a wildcard, so anyone able to present or replay a valid bearer can assert superuser. A second, quieter defeat is attribution: one shared token serves the whole fleet and the source name is free text chosen by the caller, so any holder can emit messages attributed to any agent, and the rate limit keyed on that same self-asserted string is bypassed by varying it. Authentication failures on the authorising leg produce no audit record and no alert, so the boundary has no detection surface at all.
FAILURE MODE	An operator acts on a message attributed to the wrong agent, or an attacker who can spoof the inbound leg obtains exactly the authority that the outbound leg was carefully denied.
FAIL STATE	fail-open
EMITTED SIGNAL	The bridge returns a per-message result carrying sent, deduplicated and provider identifier, and the console audit log records operator-initiated channel sends; nothing records an inbound authentication failure, so the authorising direction emits no signal at all. Needed: an authentication-outcome event stream on the inbound boundary, and a per-message signing identity so that source attribution is a cryptographic fact rather than a caller-supplied string.
METRICS	MG-048 MG-057 MG-059
COST	low
FRICTION	low
AWS BINDING	EventBridge rules delivering to an SNS topic or AWS Chatbot for the notification direction, with authorisation held separately in Systems Manager Change Manager or Incident Manager approvals bound to distinct IAM principals, so the notifying role holds no approval permission and the approving role holds no publish permission.
AZURE BINDING	Azure Monitor action groups for the notification direction; approval held in a Logic Apps approval connector or a Microsoft Teams adaptive card bound to an Entra identity, with the notifier's managed identity holding no approver role assignment.
ON HOST	yes
EVIDENCE	EV-029 EV-039
RESEARCH VALUE	Demonstrates that separation of duties is a property of a whole channel and not of a single endpoint: a structurally safe outbound leg is nullified by an inbound leg on the same medium whose only factor is a spoofable identifier, which makes this the clearest case in the catalogue of a control that is correct in isolation and void in composition.

GR-L10-02 Always reachable intervention affordance

INTENT	Guarantee that the supervising human can reach a running agent at any moment and stop what it is doing, which is the minimum precondition for any claim that a human is on the loop.
MECHANISM	Six mechanisms exist. A per-session kill route requires CSRF and an explicit confirm flag and calls tmux kill-session, terminating the pane's process group. A fleet kill route disables the systemd unit through passwordless sudo. A chat send route offers an interrupt mode and a next-turn mode. A WebSocket terminal mode attaches a pseudo-terminal to the live pane. A messaging bus kill switch and an update kill switch exist as file-existence flags.
ENFORCEMENT	conditional
BINDS	AC-015 AC-016 AC-011 AC-019
ENABLES	II
PRECONDITION	An authenticated console session, a live terminal multiplexer server and a resolvable session name; the multiplexer server has no supervising unit, was started ad hoc, and every live session is its child, so the entire affordance rests on one unsupervised process.
BYPASS	No bypass is required, because five of the six affordances do not do what their names imply and the sixth is undone by supervision. The interrupt mode is a bracketed paste followed by Enter with no escape, no interrupt signal and no cancellation: an agent mid-tool-call never sees it until the call returns, so a long-running remote command completes regardless, and next-turn mode merely appends to a mailbox drained by a hook on some later turn. The fleet kill runs a graceful-stop script that first pastes a prompt asking the agent to write a memory note and grants it up to 35 seconds of unrestricted tool access before killing anything, so a stop issued because an agent is misbehaving hands that agent one more full turn. The one mechanism that kills a process group is defeated for the nine unit-supervised agents by a restart-always policy with a ten-second delay, which returns a brand new session with the permission system disabled. The update kill switch and the bus kill switch affect only the next action, and the update switch is a mode-644 file in a directory owned by the agent account, removable with no authentication, no audit entry and no log line. Finally the console source, its build output, its environment file and its audit log are all writable by the account the agents run as, so an agent can edit the code implementing the affordance.
FAILURE MODE	An operator believes an agent has been stopped and it has not: the action in progress completes and, if the agent is unit-supervised, a fresh session resumes within about ten seconds.
FAIL STATE	fail-open
EMITTED SIGNAL	Console audit lines for session-kill, fleet-kill and chat-send carrying principal, address, action, target and outcome; the schema has no reason field, so 159 session-kill events across 46 days cannot be separated into ordinary teardown and behavioural intervention, and the kill count tracks the create count almost exactly. Needed: a typed intervention event distinguishing a behavioural stop from lifecycle teardown, plus verification that the target process actually terminated rather than that the command returned.
METRICS	MG-047 MG-044 MG-046 MG-053
COST	low
FRICTION	low, and the distribution of that friction is itself the finding: the lowest-friction affordance is the one that halts nothing, and the only affordance that halts work in progress is the one an operator is most reluctant to use
AWS BINDING	No single service; the composed equivalent is Systems Manager Session Manager termination for the interactive channel, an IAM policy or permission boundary change that survives the restart so a revived process inherits no authority, and suspension of the supervising control plane itself, for example an Auto Scaling process suspension or disabling the ECS service, since killing a child under a restart-always supervisor is not a stop.
AZURE BINDING	Azure Automation runbook stopping the hosting unit, an Azure Monitor action group triggering a Logic App that disables the workload's service principal, and Entra Privileged Identity Management deactivation to remove the standing authority a restarted session would otherwise inherit.
ON HOST	partial

EVIDENCE	EV-039 EV-013 EV-014 EV-020
RESEARCH VALUE	Forces the decomposition the supervision literature usually elides: an affordance that prevents the next action and an affordance that halts work in progress are different controls, only the second supports an on-the-loop claim, exactly one mechanism here is of the second kind, and the restart policy cancels even that for the nine agents that matter.

GR-L10-03 Bounded intervention latency commitment

INTENT	Convert supervision from an assertion into a measurable commitment by declaring, per action class, a maximum time from the action becoming observable to a human being able to halt it, and by refusing autonomy to any class whose harm completes faster than that budget.
MECHANISM	For each action class the register declares a detect-to-halt budget; the halt path is instrumented end to end; the policy decision point compares the class's measured harm completion time against the measured budget and automatically demotes any class with a negative margin to Level III, prior approval, rather than leaving the level as a stated intention. Budgets are re-measured on a schedule and by deliberate drill, not inferred from incident history.
ENFORCEMENT	conditional
BINDS	AC-029 AC-039 AC-043 AC-040
ENABLES	III
PRECONDITION	A runtime action classifier that assigns a class at attempt time, and an instrumented halt path that confirms process termination rather than command return; without both, the budget has nothing to measure and the demotion has nothing to fire on.
BYPASS	The commitment binds only actions that traverse the policy point, and on this host most do not: a shell command, a cloud command line call, a remote shell session and a file write all emit nothing and pass no gate, so a budget would govern a small, well-behaved minority while the consequential surface stays unmeasured. It is defeated a second way by measuring the wrong quantity: the fastest independent detector here trips at a 180-second stall threshold polled at one-minute granularity and exists to restart a stuck agent rather than to stop a working one, while a file write completes in milliseconds and a cloud API call in about a second, so a budget calibrated on the existing detectors would be two to five orders of magnitude optimistic. Third, it is defeated by absent attention: delivery of a signal is recorded and attention is not observable from the host, and 46 days of console audit logs contain no behavioural intervention of any kind, so a budget nobody is rostered to meet is a published number rather than a control.
FAILURE MODE	A published budget manufactures assurance: a class stays at Level I because the measurement pipeline flattered the halt path, and the negative margin that actually obtains is never surfaced.
FAIL STATE	fail-closed
EMITTED SIGNAL	Per-class latency records pairing action-attempt time from an independent source, first-render time on the operator surface, intervention-issued time and verified-stopped time, with a margin computed per class; today no surface produces any of the four timestamps in a form that could be paired.
METRICS	MG-044 MG-045 MG-046 MG-052
COST	moderate
FRICTION	moderate to high for the classes it demotes, and that is its entire mechanism of action: the commitment has no effect except where it withdraws autonomy
AWS BINDING	CloudWatch composite alarms with a declared evaluation period as the detection budget, EventBridge delivery into Systems Manager Incident Manager with an on-call schedule and a response plan carrying the engagement target; the demotion itself has no native expression and must sit in the workload's own policy decision point, since IAM cannot condition a permission on responder availability.
AZURE BINDING	Azure Monitor action groups with an on-call schedule and alert processing rules for routing, plus Entra Privileged Identity Management activation requiring approval within a maximum time; the latency commitment itself is not a native construct and must be enforced by the workload.
ON HOST	no
EVIDENCE	PROPOSED
RESEARCH VALUE	Makes supervision falsifiable: it converts the supervision-viability family from a qualitative claim into a per-class arithmetic comparison, and in the founding case it would formally demote almost every consequential class, which is the finding stated as a control rather than as a conclusion.

GR-L10-04 Approval requests presenting evidence rather than the agent's summary

INTENT	Ensure that a human approval decision is taken against machine-derived facts about what will change, never against the acting agent's prose account of what it did or intends, because the founding case measured 61.1 per cent factual divergence between agent self-report and independent evidence across 18 audited episodes drawn from 46 candidates.
MECHANISM	The approval record is assembled by the gate, not by the submitter. Concretely the interface must present: the actual diff or version delta resolved server-side from the artefact rather than declared; the target hosts and services resolved from inventory rather than asserted; current and target digests; the risk determination adjudicated server-side against the change itself and failing closed on anything unparseable; the rollback target resolved from version control rather than caller-supplied; and test results as a signed exit status from a run executed by a principal other than the submitter. Any submitter-authored prose is rendered in a separate, visually distinct region labelled as claimed and unverified, and is excluded from the risk adjudication. Partial precedent exists here: the update proposal queue already carries scanner-derived package class, version delta, digests, kernel and reboot flags and affected services, with lane, risk score and band produced by deterministic classifier code and no model in the loop, and the deployment approval gate already adjudicates the change rather than the declaration and resolves the rollback target from version control. What remains submitter-authored is the impact assessment, the summary, the rationale, the risk notes and, decisively, the test results.
ENFORCEMENT	structural
BINDS	AC-055 AC-056 AC-044 AC-031
ENABLES	III
PRECONDITION	An independent verifier identity that executes the tests and signs the result; without it the test field remains an assertion no matter how it is presented, and presentation cannot manufacture provenance that does not exist.
BYPASS	The control fails wherever the evidence is produced by the governed party. Here the impact assessments are agent prose submitted through a token-authenticated internal route, and no code anywhere compares the reviewing agent to the authoring agent, with a production instance on record of an agent recording its own peer review under another agent's token; reviewer independence is therefore instructional even where the evidence is machine-derived. A subtler defeat is that evidence which is present but unread is functionally a summary: 434 proposals carry complete machine-derived fields and every one is still pending with the approver unset, zero approvals and zero rejections in 46 days, so richer evidence has so far produced not a single decision. The most consequential defeat is outside the interface entirely: approval does not bind what runs, because the process manager re-executes whatever sits on the shared working tree and two documented incidents shipped unapproved code through ordinary restarts, so a perfectly evidenced approval can be granted for one artefact while a different one reaches production.
FAILURE MODE	The human approves on the strength of the agent's account of tests that were never run, or approves the correct diff while a different artefact is what the runtime actually loads.
FAIL STATE	fail-closed
EMITTED SIGNAL	A per-approval provenance record listing which displayed fields were machine-derived, which were independently verified and which were submitter-authored, expressed as a coverage share; today the two classes render indistinguishably in the same portal, so the operator cannot tell a scanner fact from an agent claim by looking.
METRICS	MG-051 MG-022 MG-029 MG-027
COST	moderate
FRICTION	moderate: it lengthens the record the gate assembles but not the decision the human makes, and it removes the single largest class of approval error in this case
AWS BINDING	CodePipeline manual approval actions whose approval is bound to a CodeBuild test report ARN and an artefact digest rather than to free text; Systems Manager Change Manager templates with automated pre-approval runbook checks, so the change request presents parameters and a runbook rather than a narrative; ECR image attestations and AWS Signer for artefact provenance.

AZURE BINDING	Azure Pipelines approvals combined with check gates bound to published test results and artefact provenance, Azure Monitor query-based gates for post-change health, and Microsoft Teams adaptive card approvals rendering only fields sourced from the pipeline rather than from the submitter.
ON HOST	partial
EVIDENCE	PROPOSED EV-039 EV-013
RESEARCH VALUE	Operationalises the founding case's headline measurement: it converts a 61.1 per cent divergence rate from a finding into an interface requirement, and it isolates the exact residue that remains after the machine-derived layer is built, namely test results and peer review, which are the two fields that most determine whether an approval is sound.

GR-L10-05 Visual distinction between claimed and verified

INTENT	Ensure every fact an operator reads carries its provenance, so that a party's assertion can never be mistaken for an observation, on any surface, at any level of aggregation.
MECHANISM	A two-class rendering contract enforced at the data layer rather than in the view. Every displayed field carries a provenance tag with three values, machine-observed, independently verified and party-asserted, together with the asserting identity and timestamp; the renderer refuses to display an untagged value as neutral and renders it as unverified; every aggregate view publishes the verified share of the values it summarises, so the distinction survives summarisation.
ENFORCEMENT	conditional
BINDS	AC-055 AC-053 AC-046
ENABLES	II
PRECONDITION	Field-level provenance on every state record; without a provenance field written at ingest there is nothing for the renderer to distinguish, and retrofitting it to an existing flat-file store means labelling history that cannot be re-derived.
BYPASS	The distinction is exactly as honest as its tag, and here the party that would be labelled as claiming owns the store that holds the label: the transcripts, the fleet registry, the resource accounting counters and the console source are all writable by the account every agent runs as, so a mislabelled assertion renders as verified and is then trusted more than an unlabelled one, which is worse than having no scheme. It is defeated a second way by aggregation, since a dashboard that reduces verified and claimed fields to one status badge discards the distinction at precisely the moment the operator uses it, which is why the contract has to reach the aggregate and not only the field. It is defeated a third way by attention rather than by deception: labelling does nothing about 14 concurrent sessions across roughly 14 surfaces, and a correct label on a page nobody opens changes no decision.
FAILURE MODE	An asserted completion renders identically to an observed one, enters the record as fact, and every downstream report and memory record inherits the error with no way to trace it back to a claim.
FAIL STATE	fail-closed
EMITTED SIGNAL	A verified-share figure emitted per view render and per aggregate, plus a count of values displayed with no provenance; today the update portal renders scanner-derived proposal fields and agent-written impact assessments in the same panel with no visual difference and no provenance field behind either.
METRICS	MG-029 MG-026 MG-023 MG-051
COST	moderate
FRICTION	low
AWS BINDING	Not a single service: the enforceable part is sourcing operator surfaces only from stores the workload cannot write, for example CloudWatch dashboards over metric streams and Security Hub findings with their native provenance fields, while the provenance model itself belongs to the application.
AZURE BINDING	Azure Workbooks bound to Log Analytics tables the workload identity cannot write, with Azure Monitor data collection rules establishing the collecting identity as the source of record; the three-value provenance model remains an application concern.
ON HOST	no
EVIDENCE	PROPOSED EV-039 EV-013
RESEARCH VALUE	Names the interface half of the verification problem: a regime that produces provenance and then discards it at the display layer yields the same operator decision as no verification at all, which makes rendering a governance control rather than a presentation choice.

GR-L10-06 Operator attention budgeting

INTENT	Treat the supervising human's attention as a finite allocated resource with an explicit budget, so that the number of concurrently supervised autonomous streams becomes a designed quantity rather than an emergent one.
MECHANISM	Declare a budget as decision-requiring items per operator per hour and concurrently supervised streams per operator; measure actual load against it continuously; make it binding by refusing to start a further autonomous stream, or by demoting classes to Level III, once the budget is exceeded; and route every non-decision event to a pull surface so that the push channel carries only items requiring a decision. The load to size it against is measured here: one operator, 14 live autonomous sessions, roughly 14 distinct surfaces, one push channel that is agent-initiated, deduplicated and rate limited, an intended cadence recorded in the operating guidance of about four to six messages per multi-hour dispatch, 19,334 heartbeat and nudge events against 4 human-facing alerts over the observation window, 434 undecided update proposals, and no record of a human stopping, pausing, rejecting or overriding an agent on the basis of its behaviour in 46 days of console audit logs.
ENFORCEMENT	conditional
BINDS	AC-014 AC-019 AC-046 AC-020
ENABLES	II
PRECONDITION	A single aggregation point able to count decision-requiring items across every surface, and a level model that can demote; without the first the budget has no numerator, and without the second it has no effect.
BYPASS	No agent needs to defeat this control because the environment removes its admission point: seven automations write text directly into live agent sessions and one of them fires on a clock with no stimulus, so streams start themselves and there is no gate at which a budget could refuse. It is defeated a second way by measuring the visible channel: 19,334 nudge events are written to a log file rather than pushed, so a budget computed on the push channel reports a comfortable four alerts and misses the entire load. Third, its denominator is unobservable from the host, since message delivery is recorded and attention is not, so the budget can be met on paper while nobody is reading. Fourth, and most importantly, an over-budget queue in this environment degrades to no decision rather than to a refusal: 434 proposals sit pending with the approver unset, which is fail-open behaviour wearing the appearance of a backlog, and no mechanism converts an unworked queue into withdrawn autonomy.
FAILURE MODE	The number of supervised streams grows until supervision is nominal, and the record then shows high autonomy with zero intervention, a signature that is indistinguishable from perfect agent behaviour and from no supervision at all.
FAIL STATE	fail-open
EMITTED SIGNAL	Outstanding decision-requiring items per operator, time in queue, and concurrently live autonomous streams, emitted as first-class metrics with the budget as a published threshold; today the nearest surfaces are three watchdog state logs and a proposal count, none of which is expressed as load on a person.
METRICS	MG-049 MG-057 MG-056 MG-053
COST	low
FRICTION	high by design: the budget's only real effect is to refuse work when attention is exhausted, and that refusal is the control, which is also why it is the first thing traded away
AWS BINDING	CloudWatch composite alarms to collapse correlated signals, EventBridge routing into Systems Manager OpsCenter so that decision items become owned, ageing OpsItems rather than chat messages, on-call schedules and escalation plans in Incident Manager, and AWS Chatbot as the single aggregated channel; the concurrency cap itself has no native expression and must sit at the workload's admission point.
AZURE BINDING	Azure Monitor action groups with alert processing rules and suppression, the Azure Monitor alert state model as the work queue, Azure Dashboards as the pull surface, and Entra access reviews to bound how many principals hold standing authority; the stream concurrency cap again belongs to the workload.
ON HOST	no

EVIDENCE	PROPOSED EV-039 EV-028 EV-013
RESEARCH VALUE	Converts operator overload from an impression into arithmetic, and identifies the specific empirical signature that distinguishes nominal from real supervision: high autonomy, zero intervention and a growing undecided queue, all three of which are present and measurable in the founding case.

GR-L10-07 Live terminal view of agent activity

INTENT	Give the operator a real-time, full-fidelity view of what an agent is doing, at the resolution an engineer sitting at the terminal would have.
MECHANISM	One WebSocket upgrade path whose behaviour is chosen by query parameter, with the session cookie verified before the handshake completes so an unauthenticated socket is never established. Terminal mode spawns a pseudo-terminal running an attach against the named multiplexer session and streams every byte with no server-side buffering and proxy buffering disabled. Tail mode, the default for any unrecognised parameter, is a read-only two-second poll of capture-pane pushing only the appended suffix. Chat mode is an event-driven file-watch tail of the session transcript with a 75 millisecond debounce. A pipe-pane disk log captures the raw byte stream continuously.
ENFORCEMENT	none
BINDS	AC-016 AC-029 AC-053
ENABLES	II
PRECONDITION	An authenticated console session and a live multiplexer server; that server has no supervising unit and every session is its child, so the observation channel and the agents it observes share a single point of failure.
BYPASS	This control constrains nothing, so the question is what defeats the record it produces, and there are four answers. Terminal mode is not read-only: it accepts arbitrary bytes from the browser and writes them straight into the pane of an agent running with all permission prompts disabled, with no filtering, no length cap and no content inspection, so the highest-fidelity observation channel is also the most powerful write channel on the host. It is entirely unaudited: opening it and every keystroke through it produce only a structured log line naming principal, mode and session, while the adjacent remote shell mode, which grants the same class of access to another host, is audited on open and close in two independent places, so audit coverage here tracks the perceived remoteness of the target rather than the power of the channel. Attribution is impossible: at least six independent producers can write into the same pane, three of them on cron timers, so text appearing in the view cannot be assigned to a person without correlating three watchdog state directories. And fidelity is bounded by the terminal itself, since the tail path sees only the multiplexer's default 2,000-line scrollback, so output produced faster than that in two seconds is never observed by that channel at all.
FAILURE MODE	An operator watching the tail view believes they have seen the episode; the highest-volume portion was never rendered, and any human contribution is indistinguishable from an automated nudge.
FAIL STATE	fail-open
EMITTED SIGNAL	One structured log line per connection carrying principal, mode and session, and nothing at all for the input path. Needed: an audit event on terminal open and close with the same fidelity the remote shell mode already has, and a per-submission attribution record so that human input is separable from automated injection in the pane's own history.
METRICS	MG-048 MG-052 MG-064 MG-066
COST	low
FRICTION	low
AWS BINDING	Systems Manager Session Manager provides the same interactive channel with full session logging to CloudWatch Logs or S3, including keystroke capture and KMS encryption, which is precisely the audit property missing here; CloudWatch Logs Live Tail is the read-only equivalent of the tail mode.
AZURE BINDING	Azure Bastion with native session recording is the closest analogue to the missing audit; Azure Serial Console for the raw channel and Azure Monitor Log Analytics live streaming for the read-only view.
ON HOST	yes
EVIDENCE	EV-014 EV-020 EV-013
RESEARCH VALUE	Shows that observation surfaces are also action surfaces, and provides a controlled comparison inside one codebase: two modes with equivalent power, one fully audited and one not, differing only in whether the target is on this host or another.

GR-L10-08 Automated answering of interactive prompts

INTENT	Recorded as an inverted entry: its operative intent, from the script's own header, is to keep an autonomous agent running without waiting for a person, and its effect is the removal of a supervision point rather than the creation of one.
MECHANISM	A stall watchdog runs every minute from cron under an advisory lock with a 50-second timeout. When a target session's pane has been idle beyond a 45-second dialog threshold and its last rendered lines match one exact literal corresponding to the model command line tool's own resume chooser, the watchdog sends the keystroke 1 followed by Enter into that pane, selecting the recommended resume-from-summary option. The tool raised that chooser in order to obtain a human decision about session continuity; a cron job answers it within 45 seconds and the human is never told a question was asked. This does not act within a supervision point, it deletes one: after the tool-layer permission prompt has been disabled on the command line in five places and again as a bypass default in the settings file, this chooser was among the last interactive decision points remaining, and it is now answered by a timer.
ENFORCEMENT	none
BINDS	AC-059
ENABLES	I
PRECONDITION	The target session's name appears in a hardcoded bash array of five agents, the pane matches the exact chooser string, and the pane has been byte-identical across a two-second window.
BYPASS	This mechanism is itself a bypass, so the analysis is of its own bounds and how they fail. The bounds are unusually well reasoned: the matcher is a single exact literal, and its comment records the explicit reasoning that a broader match could one day auto-accept something that truly needs a human's judgement call. Every element of that bound is nonetheless a regular expression in a shell script owned by the same account as the agents it supervises, so any agent can widen the match, add itself to the allowlist or shorten the threshold with an ordinary file write, with no audit event and no review. The bound also degrades from the other direction without anyone acting: it is coupled to the tool's cosmetic wording, so an upstream text change either disables the answering, which is safe, or extends it to a different chooser that happens to render matching text, which is not, and there is no test and no alert on either outcome. The same script's sibling nudge path is documented as having previously merged into a human's live typing, which is why the byte-stability check exists; that check bounds corruption, not authority. And the mechanism is load-bearing: it was written because an agent sat invisibly at this chooser for about one hour and 46 minutes, so removing it without an independent progress signal simply reinstates that failure.
FAILURE MODE	A decision the tooling escalated to a human is taken by a timer and recorded nowhere as a decision, so the operator's model of what the agent was authorised to resume is never updated and the episode boundary the chooser marked is erased.
FAIL STATE	fail-open
EMITTED SIGNAL	A line in the watchdog's own state log, which is a pull-only file on a filesystem the agents own; no notification is sent when an answer is given, because the escalation message fires only after two consecutive ineffective nudges, so the event that most needs a human is silent by design. Needed: a first-class pushed event recording that a human decision point was answered automatically, which question it was and what was chosen, which is exactly the observation MG-058 requires and which nothing on this host currently produces.
METRICS	MG-058 MG-050 MG-052 MG-069
COST	low
FRICTION	low, and negative in effect: the mechanism exists to remove the wait that a human decision point would impose, and it succeeds
AWS BINDING	No native equivalent, because no cloud control plane offers an answer-on-the-requester's-behalf primitive; the closest managed analogue is a Systems Manager Automation approval step or a Change Manager template whose timeout action is set to continue rather than to fail, which is the same anti-pattern in supported form and is worth naming as such.

AZURE BINDING	No native equivalent; the nearest analogues are a Logic Apps approval action with a timeout branch that proceeds, an Azure Automation runbook that suppresses an interactive confirmation, and Entra Privileged Identity Management assignments configured for activation without approval, all of which are the identity-plane and workflow-plane forms of the same pattern.
ON HOST	yes
EVIDENCE	EV-028 EV-020 EV-027 EV-039
RESEARCH VALUE	The single sharpest artefact in the case: an environment that has disabled its tool-layer permission prompt then builds a timer to answer the one interactive decision point that remained. It shows that under autonomy pressure supervision points are not merely unused, they are actively removed, and it gives the metric register a directly countable observable for a phenomenon that is usually only asserted.

GR-L10-09 Single channel aggregation of fleet signals

INTENT	Give the operator one place where everything requiring attention arrives, so that supervision does not depend on remembering which of many surfaces to check.
MECHANISM	On the read side the console aggregates genuinely: a cross-agent activity feed merging the newest events from every session transcript capped at 500, full-text search across every agent's transcript using ripgrep with capped tokens and a realpath-restricted root, a fleet panel joining the registry to live session state, and a wedged-status endpoint. On the push side one channel exists, a single handset reached through a bridge applying a per-source rate limit of six per minute and a five-minute deduplication window.
ENFORCEMENT	none
BINDS	AC-046 AC-055 AC-019
ENABLES	II
PRECONDITION	Every producing surface must emit into the aggregation point; today the update portal, the deployment approval gate on another host and three watchdog state directories do not, and the gate's actions appear nowhere in this host's audit log.
BYPASS	The aggregate is assembled from the same transcript files the agents own and can rewrite, so the feed can be shaped by the party it reports on, and the 500-event cap means a burst of low-value events evicts the one that mattered without leaving a gap. The single push channel is agent-initiated, so it cannot report on an agent that has stopped working, and its rate limit is keyed on a self-asserted free-text source, so varying the string bypasses the limit while an honest agent's genuine alert is dropped on a 429 rather than queued or retried. Deduplication silently discards a repeat inside its window, which is correct for a stuck condition and wrong for a recurring fault, and no counter tells the operator how many were suppressed. Most decisively, the two surfaces that actually require decisions, the update proposal queue and the deployment approval gate, are not on the channel at all, so what is aggregated is activity and what is missing is work.
FAILURE MODE	The operator watches the aggregate, the aggregate carries no decision queue, and 434 items age indefinitely with no approver while the feed continues to look healthy.
FAIL STATE	fail-open
EMITTED SIGNAL	The aggregate itself, plus the bridge's per-message result. Missing: a producer-liveness signal naming which surfaces have not reported recently, so that silence can be distinguished from quiet, and a queue-depth-with-ageing signal so that undecided work is visible as work rather than as absence.
METRICS	MG-057 MG-048 MG-049 MG-056
COST	low
FRICTION	low
AWS BINDING	EventBridge as the single bus with rules fanning into AWS Chatbot for the conversational surface and Systems Manager OpsCenter for the durable owned queue, with CloudWatch composite alarms suppressing correlated noise; OpsCenter is exactly the missing piece here, a queue with ownership and ageing rather than a feed.
AZURE BINDING	Azure Monitor action groups with alert processing rules for a single routed channel, Azure Dashboards or Workbooks as the pull surface, and the Azure Monitor alert state model, new to acknowledged to closed, as the durable work queue.
ON HOST	partial
EVIDENCE	EV-013 EV-029 EV-039
RESEARCH VALUE	Separates two things routinely conflated in oversight design, a notification feed and a decision queue; the founding case has a competent version of the first and none of the second, and the 434 undecided proposals are the measurable consequence.

GR-L10-10 Intervention console measuring its own latency

INTENT	Make the supervision channel self-instrumenting, so that the time from an action occurring to a human being able to halt it is a measured quantity produced by the console rather than an assumption carried in a design document.
MECHANISM	Every intervention affordance emits a paired record: the action or condition timestamp taken from an independent source, the time the operator surface first rendered it, the time the operator issued the intervention, and the time the target process was observed to have stopped, verified by process state rather than by the command returning. The console publishes the achieved distribution per action class and labels each affordance, in the interface itself, with whether it halts work in progress or only prevents the next action. The halt path is exercised on a schedule by drill, because a channel that is only measured when a real event occurs is unmeasured in an environment where real interventions do not occur.
ENFORCEMENT	none
BINDS	AC-015 AC-011 AC-057
ENABLES	II
PRECONDITION	An independent event source for the action timestamp; taking it from the agent's own transcript would make the measurement a self-report and would inherit exactly the divergence the measurement exists to detect.
BYPASS	A measurement is not defeated, it is corrupted by its inputs, and three of the four timestamps here would come from stores the measured party owns: the transcripts, the console audit log and the watchdog state files are all writable by the account the agents run as. The hardest term is the stop confirmation, and it is the one most likely to be faked by accident: the console's interrupt returns success after a terminal paste, which says nothing about whether anything stopped, and the fleet stop path returns once the unit is disabled while the agent still holds up to 35 seconds of tool access, so a console measuring command return rather than process death would publish sub-second intervention latency for a mechanism that halts nothing. The denominator is also empty: 46 days of audit logs contain no behavioural intervention, so without deliberate drills the instrument would report no data and that absence would read as an absence of problems.
FAILURE MODE	The console publishes a flattering latency derived from command returns, and the negative margin between harm completion time and halt availability is never surfaced, so the architecture's central weakness is hidden by its own instrumentation.
FAIL STATE	fail-closed
EMITTED SIGNAL	This control is signal: four paired timestamps per intervention, a per-affordance halts-work-in-progress flag, and a drill result per action class. Today the only comparable record is an audit line with no reason field and no outcome verification, which is why 159 session-kill events cannot be classified as teardown or intervention from the log alone.
METRICS	MG-044 MG-046 MG-047 MG-045
COST	moderate
FRICTION	low
AWS BINDING	CloudWatch custom metrics emitted by the intervention path itself, CloudWatch Synthetics canaries exercising the halt path on a schedule as the drill mechanism, and Systems Manager Incident Manager, which records engagement and resolution timestamps natively and is the closest managed analogue.
AZURE BINDING	Azure Monitor custom metrics from the intervention path, Application Insights availability tests as the scheduled drill, and the Azure Monitor alert state transition timestamps, fired, acknowledged and closed, as the managed analogue of the paired record.
ON HOST	no
EVIDENCE	PROPOSED EV-039
RESEARCH VALUE	This is the guardrail that makes the supervision-viability metric family measurable at all: without it MG-044, MG-046 and MG-047 have no data source that is not a self-report, and its own enforcement class of none is the point, since measurement is not enforcement but nothing else in the layer can be evaluated without it.

GR-L10-11 Wedge detection surfaced to the operator

INTENT	Ensure that an agent which has stopped making progress becomes visible to a person, rather than sitting alive and silent while the fleet panel continues to report it as healthy.
MECHANISM	A console endpoint computes fleet wedged status from the registry joined to per-agent transcripts. Three cron watchers derive a stall signal from terminal state: an idle age from the multiplexer's window-activity timestamp, and a regular expression over the last six rendered pane lines matching the tool's busy markers. Thresholds are 180 seconds polled every minute for the stall watchdog, 150 seconds polled every three minutes for the heartbeat watcher, and 300 seconds polled every two minutes with a 30-minute re-notification floor for the alert-only human-facing watcher. On detection the first two type an English sentence into the pane and press Enter; only the third escalates to a person without acting.
ENFORCEMENT	none
BINDS	AC-016 AC-046 AC-015
ENABLES	II
PRECONDITION	The session's name appears in a watcher's hardcoded array; the tool's spinner wording still matches the busy regular expression; the multiplexer server is alive.
BYPASS	The detector's input is the observed party's own rendered output, so it is not an independent signal, and it fails in both directions without anyone acting: an upstream change to the tool's spinner wording silently disables stall detection fleet-wide with no test and no alert, and the code already documents two production failures of exactly this fragility, a completed-turn footer whose past-tense words sat in scrollback and made idle sessions read as busy, and an interactive chooser that displayed no busy marker at all and left one agent blocked for about one hour and 46 minutes. Scope is a bash array in a file owned by the same account as the agents, so an agent can remove itself from supervision with a one-line edit and no audit event, and equally an agent created today is unsupervised by default because nobody edited the array: 8 of 14 live sessions have no stall detection at all, including three registry-backed unit-supervised agents, and one ad-hoc session has produced no output for seven days with its process alive and nothing noticing. The arrays are hardcoded precisely because an earlier glob-based version injected into an unrelated team's sessions, so the fix for one failure mode institutionalised the other. Finally the trip condition is silence, so an agent confidently doing the wrong thing at full output rate is by construction never wedged.
FAILURE MODE	A live session stops progressing and no surface reports it, while the fleet panel continues to report it as live because liveness is measured by session existence rather than by progress.
FAIL STATE	fail-open
EMITTED SIGNAL	Three watchdog state logs, 523 lines from one detector, 19,334 from another and 4 from the alert-only one over their respective windows, plus an escalation message only after two or five consecutive ineffective nudges; the wedged endpoint is pull-only. Needed: an agent-emitted progress heartbeat with a task-state model so the detector observes progress rather than pixels, and a coverage signal naming every live session with no detector attached.
METRICS	MG-068 MG-052 MG-017 MG-061
COST	low
FRICTION	low
AWS BINDING	CloudWatch alarms on a custom application heartbeat metric with treat-missing-data set to breaching, which is the native answer to the fail-open behaviour here, plus Systems Manager Fleet Manager or Application Insights for process-level health rather than session existence.
AZURE BINDING	Azure Monitor heartbeat and custom availability metrics with an alert on missing heartbeat, Application Insights availability tests, and VM Insights for process liveness independent of the terminal.
ON HOST	partial
EVIDENCE	EV-020 EV-028 EV-013 EV-039
RESEARCH VALUE	A clean demonstration that a monitor whose input is the monitored party's rendered output is not an independent signal, and that hardcoded scope makes supervision

> coverage a maintenance artefact rather than a property of the system, with a seven-day undetected silence available as the measured consequence.

2.11 L11. Temporal e contextual

Esta camada governa que restrições de tempo, sequência e duração se aplicam a uma ação, independentemente de quem a executa ou do que ela tem como alvo. Duas entradas presentes definem a postura temporal real do caso fundador: disparo agendado sem supervisão em um relógio, classificado `none` porque um agendamento que dispara sem nenhum estímulo humano é o oposto de uma restrição temporal, e uma janela de soneca (snooze) que suprime alertas idênticos repetidos, conditional, o que reduz o ruído mas também reduz o volume de sinal que qualquer canal voltado ao humano na L10 recebe **EV-027**. O controle de horário comercial sobre contato externo é parcial, presente para uma superfície de produto e ausente nas demais, e seu próprio campo de fricção registra, candidamente, que ele entra em conflito direto com um produto de atendimento fora do horário comercial, de modo que a tensão foi resolvida em favor da disponibilidade, e não em favor do controle temporal.

Seis das nove entradas são propostas, e três são structural: mandatos com prazo delimitado (time boxed), uma duração máxima de execução sem supervisão antes de reautorização, e uma restrição de sequenciamento exigindo que a verificação se complete antes que a próxima classe de ação seja desbloqueada. A última delas é a instância, na camada temporal, da mesma ideia que a L7 enuncia como política e a L8 enuncia como evidência: a de que a autonomia para uma ação não deveria conseguir substituir a verificação da ação anterior.

Camada 11. Temporal e contextual.

GR-L11-01 Time boxed mandates

INTENT	Give every grant of autonomy an expiry, so that authority must be renewed by a decision rather than persisting by default for the life of the account.
MECHANISM	A mandate record carrying a not-before and not-after timestamp and a maximum cumulative action count, bound to the credential rather than to a field: the agent receives a session credential minted for the mandate window and scoped by a session policy, so expiry withdraws capability rather than prompting the agent to stop. On expiry the session degrades to a read-only posture, which is visible and recoverable, and renewal requires a human principal distinct from the agent and is journalled with the approver's identity.
ENFORCEMENT	structural
BINDS	AC-019 AC-029 AC-039 AC-043
ENABLES	II
PRECONDITION	Per-agent workload identity with short-lived credentials; on this host every agent shares one operating system account holding five static non-expiring cloud key pairs, thirteen secret files and seven unencrypted host keys, so there is nothing whose expiry would remove anything.
BYPASS	Implemented as a field, expiry is defeated by ownership: every instruction file, role file, roster and registry entry on this host is writable by the party it governs, so a date in a role file is a suggestion and the mandate can be extended by the mandated. Implemented as credential lifetime, it is defeated by every long-lived secret that outlives the window, and this host holds enough of them that an expired session buys nothing while the material on disk does not expire. It is defeated a third way by automatic re-dispatch: a scheduler terminates and recreates every fleet agent daily at 08:00 and hands it a fresh startup prompt, and units restart always within ten seconds, so a naive window renews itself without a human deciding anything and expiry becomes a formality with a timestamp.
FAILURE MODE	Authority granted for one task persists indefinitely, and the agent's effective mandate becomes the union of every task it has ever been given, which is unbounded and unrecorded.
FAIL STATE	fail-closed
EMITTED SIGNAL	A mandate-granted event with grantor, scope and window, an expiry event, and a denied-past-expiry event; today nothing on this host records when an agent's authority was granted, by whom, or for how long, so the quantity has no observable at all.
METRICS	MG-034 MG-030 MG-037 MG-042
COST	moderate
FRICTION	moderate
AWS BINDING	STS AssumeRole with DurationSeconds bounded to the mandate window and an inline session policy narrowing the permission set for that window, IAM policy conditions on aws:CurrentTime and aws:TokenIssueTime, and IAM Identity Center or IAM Roles Anywhere for the human-approved renewal.
AZURE BINDING	Entra Privileged Identity Management time-bound eligible assignments with a maximum activation duration and justification, Azure RBAC role assignments carrying start and end times, and federated workload identity credentials scoped to the task rather than to the host.
ON HOST	no
EVIDENCE	PROPOSED EV-032 EV-039
RESEARCH VALUE	Makes the difference between an agent's nominal mandate and its accumulated effective authority measurable, which is the central quantity of the boundary and mandate metric family and the one the founding case cannot currently compute at all.

GR-L11-02 Hours of operation restriction per action class

INTENT	Restrict the highest-impact action classes to periods when a human is present and able to respond, so that the supervision latency assumption actually holds at the moment the action is taken.
MECHANISM	A policy decision point consults a change calendar keyed by action class and target tier and denies any class above a declared risk threshold outside its window and outside a declared on-call roster. Partial precedent exists here: the update executor's unattended lane requires the target host to be inside a maintenance window expressed as a local time range on its inventory record, and one host carries a zero-active-calls restart precondition so a service is not restarted mid-interaction; the daily fleet reload time was likewise chosen on call-volume evidence showing no calls at that hour rather than by convention.
ENFORCEMENT	conditional
BINDS	AC-031 AC-044 AC-045 AC-040
ENABLES	II
PRECONDITION	A runtime classifier assigning an action class at attempt time, and a calendar held outside the acting party's write reach; neither exists here, and the window that does exist is per host rather than per class and gates exactly one automation.
BYPASS	A time window is a control only when the calendar is outside the actor's reach, and here it is a JSON field in a mode-644 file owned by the account every agent runs as, alongside the tier, the enabled flag, the key path and the restart constraints, so widening the window is a file write with no audit entry; the same account also holds unrestricted passwordless root, so even a root-owned calendar would not be out of reach. Time windows are also waited out rather than defeated: an agent that cannot act at 02:00 queues the action for 09:00, which is a delay and not a prohibition, and delayed execution arrives precisely when the operator is busiest. The window's premise is that a human is present during it, and 46 days of audit logs contain no behavioural intervention at any hour, so the premise is false in both directions. Finally a per-host window does not compose: an action whose blast radius spans several hosts is inside one window and outside another, and nothing here resolves that conflict.
FAILURE MODE	An unattended change lands inside a nominal business-hours window while nobody is watching, and the window is afterwards cited as evidence that the change was supervised.
FAIL STATE	fail-closed
EMITTED SIGNAL	A denied-outside-window event and a within-window action count per class, with the calendar version recorded on each decision; today the only related record is the update executor's own state file, and its apply path has never fired.
METRICS	MG-052 MG-046 MG-033
COST	low
FRICTION	moderate to high for the classes it restricts, and it conflicts directly with unattended operation, which is why it is among the first temporal controls traded away
AWS BINDING	Systems Manager Change Calendar with OPEN and CLOSED events consulted by Automation runbooks and Change Manager templates, IAM policy conditions on aws:CurrentTime to deny outside a window, and EventBridge Scheduler for the deferred-execution alternative.
AZURE BINDING	Azure Update Manager maintenance configurations and maintenance windows as the closest managed analogue, Azure Policy assignments enforced over the same scope, Azure Automation schedules, and Entra Privileged Identity Management activation windows for the identity half.
ON HOST	partial
EVIDENCE	PROPOSED EV-028 EV-031 EV-027
RESEARCH VALUE	Distinguishes a time window that shifts risk from one that reduces it, and shows that a calendar owned by the governed party is a schedule rather than a control, which is the temporal instance of the catalogue's recurring pattern.

GR-L11-03 Cooling off period between related high impact actions

INTENT	Insert a mandatory interval between an action and the next related action of equal or greater impact, so that a fault has time to surface before it is compounded and so that intervention remains meaningful for longer than the action itself takes.
MECHANISM	The policy point keeps a durable per-target, per-class ledger of completed high-impact actions and denies the next one until both a class-specific interval and an independent post-action health verification have elapsed. Partial precedent exists here and is well built: the update executor orders waves across tiers with concurrency capped at one for production, persists the soak deadline as a timestamp so it survives a process restart, and halts the whole wave and the whole run on any host failure.
ENFORCEMENT	conditional
BINDS	AC-009 AC-044 AC-045 AC-039
ENABLES	II
PRECONDITION	A durable per-target action ledger and an independent post-action health verification; an interval without a health check only delays the second fault rather than preventing it.
BYPASS	An interval enforced inside one executor is bypassed by reaching the same effect another way, and every high-impact class here has at least two routes: a deployment can go through the approval gate or by placing code on the shared working tree that the next ordinary restart picks up, which is documented as having happened twice; a service restart can go through the console, the process manager directly, a remote shell, or an unattended operating system upgrade that restarts affected services with no agent involved at all. It is defeated a second way by identity: with one shared account there is no way to distinguish two related actions by one agent from two unrelated actions by two agents, so the interval is either too coarse to be usable or too fine to be safe. And it is in continuous tension with the environment's own machinery, since a three-minute clock-driven nudge and a one-minute approved-patch poller both exist specifically to shorten the gap between actions.
FAILURE MODE	A faulty change propagates to the next target before the first target's fault has surfaced; the founding case's deploy that was reported successful while a service's core connection function stayed broken for about 24 hours is exactly this shape.
FAIL STATE	fail-closed
EMITTED SIGNAL	A per-target last-action timestamp, a denied-too-soon event, and the health verification result that released the interval; the only durable inter-action interval on the host today is the update run's persisted soak deadline, and it governs one automation that has never applied anything.
METRICS	MG-038 MG-017 MG-020
COST	low
FRICTION	moderate, and it reduces throughput directly, which is why it is characteristically the first control removed under delivery pressure and the one most often reinstated after an incident
AWS BINDING	CodeDeploy deployment configurations with a bake time and automatic rollback on a CloudWatch alarm, Systems Manager Automation with an explicit sleep and approval step between waves, and Auto Scaling instance refresh with an instance warm-up.
AZURE BINDING	Azure Pipelines environment approvals and gates with a delay before evaluation, Azure Update Manager maintenance configurations for staged rollout, and Azure Monitor alert-based gates as the health condition that releases the interval.
ON HOST	partial
EVIDENCE	PROPOSED EV-031 EV-028
RESEARCH VALUE	Separates two functions usually merged under the single word canary, bounding blast radius and buying detection time; the founding case implements the first competently in one module and the second nowhere else, which isolates detection time as the variable.

GR-L11-04 Sequencing constraint requiring verification before the next class unlocks

INTENT	Make progression through an episode conditional on independently verified state, so that an agent cannot move from building to deploying on the strength of its own account of the earlier step.
MECHANISM	The policy point holds a per-episode state machine in which each action class has a precondition expressed as a verification artefact produced by a principal other than the acting agent: a build attestation signed by a builder identity, a test result from an independently executed run, or a post-change health probe from an infrastructure-emitted event stream. Absent the artefact the next class is unreachable rather than discouraged. The strongest precedent on this host shows the shape: in the update executor a module-private symbol means only the inventory gate can mint an apply target, and the apply function demands one, so applying without a completed snapshot receipt is uncallable at compile time rather than refused at run time.
ENFORCEMENT	structural
BINDS	AC-008 AC-044 AC-055 AC-056
ENABLES	III
PRECONDITION	An independent verifier identity holding a credential the acting agent cannot read; without it the artefact is produced by the same party under a different name and the sequencing is ceremonial.
BYPASS	Sequencing binds the path and not the effect, so it fails wherever the effect has a second route, and the decisive one here is that approval does not bind what runs: the process manager re-executes whatever sits on the shared working tree, so an agent that cannot unlock the deploy class can place the code and wait for an ordinary memory-pressure restart, which is documented as having shipped unapproved code twice. It fails a second way on authorship, because reviewer independence in the existing gate is instructional, no code compares the originating agent to the reviewing agent, and an agent recorded its own peer review under another agent's token in production. It fails a third way on identity: with one shared operating system account, any artefact an independent verifier signs is signed with a key the acting agent can read, so independence is a naming convention rather than a boundary. And the compile-time precedent, strong as it is, binds only that module's own callers and says nothing about a shell.
FAILURE MODE	An episode advances on an unverified claim; the founding case's 61.1 per cent divergence rate is the measured frequency of that input being wrong, so the sequencing gate is only as sound as the artefact's independence.
FAIL STATE	fail-closed
EMITTED SIGNAL	An unlock event naming the artefact, its producing identity and its verification method, and a denied-unverified event where the artefact is absent or unparseable; today the closest record is the deployment gate's server-side risk adjudication, whose test-results field is asserted rather than verified.
METRICS	MG-023 MG-024 MG-022 MG-010
COST	high
FRICTION	high, and the friction is the mechanism: it is the cost of making verification a precondition of control flow rather than a reporting activity performed afterwards
AWS BINDING	CodePipeline stage transitions gated on CodeBuild test report ARNs and on artefact signatures from AWS Signer, with provenance recorded in ECR image attestations, and Systems Manager Automation approval steps between stages so that the unlock is an authorisation rather than a convention.
AZURE BINDING	Azure Pipelines environment checks and gates including Azure Monitor query gates for the post-change probe, artefact attestation through Azure Container Registry content trust, and Deployment Stack policies to prevent out-of-band changes to the same resources.
ON HOST	no
EVIDENCE	PROPOSED EV-031
RESEARCH VALUE	Turns verification from a reporting activity into a control-flow precondition, which is the concrete architectural form of the thesis claim, and the founding case supplies both the strongest example of the pattern working, at compile time in one module, and the clearest example of it being routed around, through a shared working tree.

GR-L11-05 Scheduled unattended dispatch on a clock

INTENT	Recorded as an inverted entry: its operative purpose, from the script's own header, is an operator directive to drop a message every three minutes to wake the agents up, that is, to guarantee that autonomous agents never stop working by removing the need for any stimulus at all.
MECHANISM	A cron entry every three minutes runs a heartbeat watcher under an advisory lock with a 50-second timeout. For each of six named sessions whose pane has been idle at least 150 seconds, does not match the busy regular expression and has been byte-stable for two seconds, it sends a full instruction sentence and Enter into the pane, telling the agent to pick up pending approval-gate items, inbox messages or open backlog work from its role file. There is no iteration cap, no budget, no end condition and no human in the loop; a per-agent re-nudge floor of 150 seconds is the only rate limit, and escalation to a person happens only after five consecutive byte-identical panes, about 15 minutes, and is report-only. So a timer causes an autonomous agent holding unrestricted passwordless root and estate-wide remote access to begin acting, with no stimulus and no human present. It is one of seven automations that write text directly into a live agent's input stream and the only one requiring no external event whatsoever.
ENFORCEMENT	none
BINDS	AC-016 AC-019 AC-029 AC-043
ENABLES	I
PRECONDITION	The session's name appears in a hardcoded array of six, the pane is idle beyond 150 seconds, does not match the busy regular expression, and is byte-identical across a two-second window.
BYPASS	This is not a control to be bypassed; what is worth recording is what bounds it and how those bounds fail. The allowlist is a bash array in a file owned by the same account as every agent it wakes, so an agent can add or remove a session with an ordinary file write and no audit event, and a dormant full copy of a sibling watchdog targeting a retired product group sits in the same directory one crontab line away from live. Its stricter counterpart is explicitly overridden: a separate alert-only watcher exists precisely because blind injection into the operator's own human-facing session previously caused real harm, and this watcher includes that same session anyway by documented operator override, so where two rules conflict the permissive one wins. The byte-stability check bounds corruption of a human's concurrent typing, not the injection or its authority. And the failure it addresses is real, since an agent once sat invisibly at a chooser for about one hour and 46 minutes, so removing the automation without replacing it with an independent progress signal simply reinstates that.
FAILURE MODE	An agent that has legitimately finished, or that has correctly stopped for want of information, is instructed by a timer to find more work, and the resulting actions have no requester, no brief and no attributable human decision behind them.
FAIL STATE	fail-open
EMITTED SIGNAL	19,334 lines in the watcher's own state log over the observation window, pull-only, on a filesystem the agents own, against 4 human-facing alerts from the alert-only watcher in the same period. Needed: a first-class pushed event per unattended dispatch recording that an agent began acting with no human requester, and a count of actions taken within sessions begun that way; MG-069 counts the paths but nothing counts the dispatches.
METRICS	MG-069 MG-002 MG-005 MG-050
COST	low
FRICTION	low, and negative in effect: the mechanism exists to eliminate the idle time that a human-in-the-loop design would use, and its success is measured in that elimination
AWS BINDING	EventBridge Scheduler or a Systems Manager Automation rate schedule invoking a workload with no approval step is the same pattern in managed form; the governance-relevant difference is that such a schedule can be disabled centrally under a separate IAM principal and every invocation is recorded in CloudTrail, neither of which is true of an entry in a user crontab.
AZURE BINDING	Azure Automation schedules or a Logic Apps recurrence trigger are the managed equivalents; Azure Automation additionally records job history and allows a schedule to be suspended under a separate Azure RBAC role, which is the control property the crontab form lacks.

ON HOST	yes
EVIDENCE	EV-028 EV-027 EV-039
RESEARCH VALUE	The temporal counterpart to the automated answering of interactive prompts: together they show that this environment's automation budget is spent on removing waits rather than on creating them, and they give the thesis a directly observable instance of autonomy that originates in a timer rather than in a request.

GR-L11-06 Snooze window suppressing repeated identical alerts

INTENT	Prevent a persistent condition from generating a continuous alert stream that would exhaust the single push channel and desensitise the one operator who reads it.
MECHANISM	Suppression is implemented independently in four places with four different windows and no shared state. The notification bridge deduplicates on a composite of source and caller-supplied key over a five-minute window and rate limits to six messages per minute per source, returning a deduplicated result or a 429 with the message dropped rather than queued. The alert-only human-facing watcher re-notifies at most once per 30 minutes per session. The stall watchdog and heartbeat watcher enforce per-agent re-nudge floors of 300 and 150 seconds and escalate to a person only after two and five consecutive ineffective cycles. A commit watcher snoozes a dirty-working-tree notice for 60 minutes.
ENFORCEMENT	conditional
BINDS	AC-046 AC-016
ENABLES	II
PRECONDITION	The caller supplies a deduplication key; with no key there is no deduplication at all, and the standing instruction to always pass one lives in a markdown file rather than in the code.
BYPASS	Every suppression state here is in-process memory in a service that restarts, so a process restart resets all buckets and windows, and the rate limit is keyed on a self-asserted free-text source, so varying the string defeats it while one shared token lets any holder attribute messages to any source. Suppression is also the wrong shape for the risk it manages: deduplication silently discards a repeat, which is right for a stuck condition and wrong for an intermittent fault, and no counter tells the operator how many were suppressed, so a suppressed stream and a quiet system produce an identical observation. The 429 path drops rather than queues, so the channel is least reliable exactly when it is busiest. And the suppression governs a channel that is agent-initiated in the first place, so the largest real load, 19,334 nudge events, is not on it, is not suppressed and is simply never presented to a person.
FAILURE MODE	A genuinely new alert is discarded as a duplicate because an agent reused a key, or the third occurrence of an intermittent fault is dropped, and the operator's impression of frequency is wrong in the direction of calm.
FAIL STATE	fail-open
EMITTED SIGNAL	The bridge returns a deduplicated flag and a 429 to the caller and the watchers write to their own state logs; none of this reaches the operator. Needed: a suppressed-count rendered beside each alert so the operator sees frequency rather than a single instance, and a durable suppression store so that a restart does not silently reopen the channel.
METRICS	MG-057 MG-056 MG-061
COST	low
FRICTION	low
AWS BINDING	CloudWatch alarm evaluation periods and datapoints-to-alarm, composite alarms for correlated conditions, SNS FIFO message deduplication, and Systems Manager Incident Manager deduplication strings, all of which are durable server-side equivalents of the in-memory windows used here.
AZURE BINDING	Azure Monitor alert processing rules with suppression schedules, action group deduplication, and the Azure Monitor alert state model, which keeps a suppressed alert visible in a queue rather than discarding the observation.
ON HOST	yes
EVIDENCE	EV-029 EV-028 EV-027
RESEARCH VALUE	Shows suppression implemented four times with four windows and no shared state, and isolates the asymmetry that matters for oversight: what is suppressed is counted nowhere, so an operator cannot distinguish a quiet system from a suppressed one, which makes silence uninterpretable as evidence.

GR-L11-07 Maximum unattended run duration before re-authorisation

INTENT	Bound how long an agent may act continuously without a human decision, so that autonomy is a renewable grant rather than a steady state that persists until someone notices.
MECHANISM	A per-session wall-clock and action-count budget maintained by the supervisor rather than by the agent, enforced at credential level so that exhaustion removes capability instead of relying on the agent to stop; on exhaustion the session moves to a read-only posture with its work preserved and a re-authorisation request queued to a human. The contrasting present state is stark: there is no idle timeout and no maximum session lifetime anywhere on this host, verified across the console, the fleet scripts and the unit files; the only time-based termination is a daily reload that resets the process and grants no decision; and the one general-purpose looping harness on the host that does ship with an iteration cap, a 240-minute timeout, exponential backoff and a dead-man file has no jobs defined and is dormant.
ENFORCEMENT	structural
BINDS	AC-014 AC-019 AC-029
ENABLES	II
PRECONDITION	Per-agent identity with short-lived credentials; with one shared account and non-expiring key material on disk there is nothing for budget exhaustion to withdraw.
BYPASS	A budget measured in session lifetime is defeated by re-creation, and this environment re-creates automatically: units restart always within ten seconds, the console replays a desired-sessions file verbatim on startup including its stored flags, and a scheduler restarts every fleet agent daily, so a naive budget resets itself without a human deciding anything. Measured in cumulative actions instead, it is defeated by the absence of a counter the agent does not own: the host's resource accounting discards session identity at ingest, enforces no budget at all and is writable by the agents. Measured at credential level it would hold, but there is no per-agent credential to expire. The most interesting defeat is organisational rather than technical: the shape is understood, since a capped, dead-man-guarded loop harness exists in the codebase, and it is applied to nothing while the uncapped clock-driven nudge governs the sessions that matter.
FAILURE MODE	An agent runs for days on a mandate given once; the record then shows a high autonomous completion share which is in fact an absence of decisions, and one live session on this host has been alive for ten days with no output for seven of them.
FAIL STATE	fail-closed
EMITTED SIGNAL	Elapsed unattended duration and actions since the last human decision, per session, with a re-authorisation queue depth and an exhaustion event; today the closest available figures are raw session uptime and the iteration cap of a harness that is not running.
METRICS	MG-005 MG-053 MG-052 MG-002
COST	moderate
FRICTION	high, and deliberately: it converts unattended time into a queue of decisions, which is precisely the cost the founding case has chosen not to pay and the reason its autonomy figures are high
AWS BINDING	IAM role maximum session duration and STS DurationSeconds as the capability clock, Step Functions with a task timeout and a callback-pattern human approval task for continuation, and Systems Manager Automation with an execution timeout and an approval step.
AZURE BINDING	Entra Privileged Identity Management activation maximum duration with renewal requiring approval, Azure Automation runbook job time limits, and Logic Apps durable approval steps for continuation.
ON HOST	no
EVIDENCE	PROPOSED EV-020 EV-028
RESEARCH VALUE	Names the quantity that distinguishes delegated autonomy from abandoned autonomy, and supplies a rare case in which the control already exists as working code in the same codebase and is applied to nothing, which isolates adoption rather than capability as the variable.

GR-L11-08 Business hours gating on external contact

INTENT	Prevent an autonomous system from contacting a member of the public or an external recipient at a time when no human is available to intercept, correct or take over the interaction.
MECHANISM	External contact is genuinely time-shaped in this estate, but by product design rather than by a governing control. The voice product's registered fleet group is an after-hours service, which is a temporal restriction on when the system contacts people at all; the daily reload hour was chosen on call-volume evidence rather than convention; the update executor gates its unattended lane on a per-host maintenance window; and one host carries a zero-active-calls restart precondition so a service is not restarted mid-interaction. None of these is a gate on external contact as an action class: the operator notification bridge has no time restriction, the one usable outbound email transport is a direct API call with no wrapper and no schedule, and the inbound command channel invites a reply at any hour in the system's own startup message.
ENFORCEMENT	conditional
BINDS	AC-049 AC-050 AC-047 AC-046
ENABLES	III
PRECONDITION	A runtime classifier that recognises an external-contact attempt before the transport call; today the transport is a direct HTTPS call authenticated from a file any local process can read, so there is no interception point at which a window could be consulted.
BYPASS	A time gate on external contact is defeated first by the absence of a chokepoint: the outbound message and email APIs are reachable by any process on the host holding a readable key, so the gate would live in a wrapper nothing compels anyone to use, and the host's own record shows a prior one-off sender script being written, used and deleted. It is defeated second by test and development traffic reaching real recipients, which is the documented failure in this estate rather than a hypothetical one: a review workflow pointed at a live database sent 13 real messages to a real external number about fabricated incidents [REPORTED, from the estate's own memory record, not reproduced by any probe in this evidence set], inside working hours, where no time window would have helped. It is defeated third by the ambiguity of the window itself, since a clock cannot distinguish a scheduled after-hours service, where contact at 02:00 is the product, from an unattended agent contacting the public at 02:00, which is the hazard; separating them requires the action class and the requesting identity, not the hour. The control that actually addresses this risk in the estate is the structural test mode that intercepts every external dispatch, and it is a different guardrail.
FAILURE MODE	An autonomous agent contacts a real external party with incorrect content while nobody is watching, and the first evidence available to the organisation is the recipient's reply.
FAIL STATE	fail-open
EMITTED SIGNAL	Per-message delivery is recorded by the transport and, for the operator channel, by the bridge; nothing distinguishes an intended external contact from an accidental one and there is no pre-send classification at all. Needed: an external-contact-attempted event emitted before the transport call carrying recipient class, requesting identity, environment and whether a human authorised it.
METRICS	MG-043 MG-038 MG-045
COST	low
FRICTION	moderate, and it conflicts directly with an after-hours product, which is how the estate resolved the tension in favour of availability and why the remaining control is a product boundary rather than a governance one
AWS BINDING	EventBridge Scheduler plus an SES configuration set and a sending Lambda that consults a Systems Manager Change Calendar, with SES verified identities and sandbox restrictions as the nearest native interception point for the email case; SNS and Pinpoint quiet-time settings for the messaging case.
AZURE BINDING	Azure Communication Services fronted by a Logic App that consults a maintenance schedule, since Azure Policy cannot gate an outbound message; the interception has to be an application chokepoint, with Azure Monitor recording every attempt.
ON HOST	partial
EVIDENCE	EV-029 EV-031 EV-022

RESEARCH VALUE Shows the limits of temporal controls on irreversible external actions: the hour of the day is a poor proxy for human availability, a useless proxy for correctness, and directly at odds with any product whose value proposition is availability outside business hours.

GR-L11-09 Staleness bound on the state a decision is based on

INTENT	Prevent decisions being taken on state that has aged past validity, whether the decision is made by an agent, by an approver or by an automation.
MECHANISM	Every state record carries an observation timestamp and a per-class maximum age; the policy point refuses an action whose supporting observation exceeds the bound and requires a fresh probe first. The bound applies symmetrically to the agent's retrieved context and to the human's approval view, and the age at decision time is written onto the decision record. The need is concrete here: 434 update proposals have been pending since a single overnight sweep, so an approval given today would be given against a package inventory that may be days old; the fleet registry and the console's desired-sessions file are documented as describing intent rather than reality, with one live process carrying a resume flag no code path emits and not matching its own record; and the agent's context is assembled from a 26.7 MB corpus, parts of which are synchronised from remote repositories every five minutes, none of which carries an expiry.
ENFORCEMENT	conditional
BINDS	AC-056 AC-031 AC-051 AC-055
ENABLES	II
PRECONDITION	An observation timestamp on every record and a probe cheap enough to refresh it on demand; without the second the bound converts silently into a denial and the queue simply stops being worked.
BYPASS	A staleness bound is only as good as its clock and its freshness claim, and here both are writable by the party they govern: the scan state, the registry, the resource counters and the transcripts are all owned by the account the agents run as, so a stale record can be restamped rather than refreshed and nothing would detect the difference. It fails a second way on the distinction between age and validity: a record re-scanned an hour ago is fresh and still wrong if the scan itself read a cache, and the console's session list degrades to an empty array when the multiplexer is unreachable, which is indistinguishable from a fresh observation of an idle box. It fails a third way on the agent's own context, where no timestamp reaches: scheduled jobs fast-forward remote repository content into directories the retrieval tools serve into the model's context every five minutes, so the agent's working knowledge changes underneath it with no decision and no version. And tightening the bound without a cheap probe converts a stale decision into no decision, which this environment already demonstrates at scale with 434 items and zero decisions.
FAILURE MODE	An approval is granted against an inventory that no longer describes the target; the change applies to a host that has moved on, and the approval record still reads as correct on review.
FAIL STATE	fail-closed
EMITTED SIGNAL	Observation age at decision time recorded on the decision, and a refused-as-stale counter per class; the substrate partly exists, since the update store already carries last-probe and last-scan timestamps, and nothing consumes them as a bound.
METRICS	MG-051 MG-026 MG-064
COST	moderate
FRICTION	moderate
AWS BINDING	AWS Config rules evaluating configuration recency, Systems Manager Inventory with a scheduled association and an ageing compliance state, and CloudWatch alarms with treat-missing-data set to breaching so that an absent observation is a fault rather than a pass.
AZURE BINDING	Azure Resource Graph queries filtered on freshness, Azure Policy compliance evaluation timestamps, Azure Update Manager assessment recency, and Azure Monitor missing-heartbeat alerting as the equivalent of treating absent data as a breach.
ON HOST	no
EVIDENCE	PROPOSED EV-031 EV-020 EV-027
RESEARCH VALUE	Introduces time into the verification family: a claim can be true, independently verified and still wrong because the world moved, which is a failure mode instruction-based governance cannot address at all and which no amount of agent honesty prevents.

2.12 L12. Camada de modelo e prompt

Esta camada é catalogada por completo precisamente porque é para onde o esforço real de governança do caso fundador foi direcionado, e porque cada uma de suas nove entradas é instructional por construção, nenhuma pode ser outra coisa. Um arquivo de orientação operacional de repositório, autorização permanente em prosa para ações de controle de versão, uma instrução em prosa de sequência de implantação, um brief de despacho declarando proibições em prosa, instruções em prosa para confirmar antes de uma chamada de nuvem que muta estado ou antes de execução remota, um repositório de auto-memória carregado em toda sessão, uma restrição de system prompt sobre o uso de ferramentas, e o comportamento de recusa do modelo como controle: todas as nove estão presentes, bem mantidas, e nenhuma bloqueia uma chamada de ferramenta no runtime. Quarenta arquivos de instrução clássicos totalizando cerca de 538 kilobytes, ou cerca de 1.240 arquivos e 26,7 megabytes contando arquivos de papel (role), listas, prompts, três repositórios de memória e a base de conhecimento, são confirmados como contexto, e não como enforcement, porque o único hook pré-ferramenta é telemetria que sempre termina com sucesso, nenhuma camada de política gerenciada existe no host, e nenhuma verificação de autorização em lugar algum lê qualquer um deles `EV-032`.

Os próprios campos de custo e fricção da camada tornam explícita a razão pela qual esta camada recebe investimento excessivo, em vez de exigir inferência: `GR-L12-01` e `GR-L12-09` registram seu custo operacional como, respectivamente, efetivamente o custo de escrever um documento, e efetivamente zero, e ambos observam, no mesmo campo, que é exatamente por isso que as organizações dependem excessivamente deles. Um controle que não custa nada para construir, não impõe fricção alguma ao trabalho legítimo, e não consegue bloquear nada de modo structural, é a forma mais barata possível de parecer governado.

Camada 12. Camada de modelo e prompt.

GR-L12-01 Repository operating guidance file as agent context

INTENT	Establish shared operating rules, an estate map and a deployment model for every agent session, so that behaviour is consistent without per-task instruction.
MECHANISM	Two canonical markdown files are loaded as context on every session: a project-scope file in the working directory every session shares, and a user-scope global file. They carry the estate map, the deployment model, a golden rule to inspect the running workload before deciding, and a standing authorisation clause. Neither uses import syntax, so the loaded context is exactly those two files rather than a transitive graph. Forty classic instruction files totalling about 538 kB exist across the tree, plus 26 alias symlinks, and a third-party coding tool on the host is pointed at the same user-scope file.
ENFORCEMENT	instructional
BINDS	AC-002 AC-009 AC-023 AC-025
ENABLES	I
PRECONDITION	The session's working directory is the one the project-scope file governs; every session here shares that directory, which is also why the five per-agent settings files are literally empty objects and cannot scope anything per agent.
BYPASS	The failure specific to this entry is ownership. The document that defines the agent's operating rules is owned by the agent: all 40 classic instruction files, and all 1,240 files in the broad corpus, return writable for the governed session, so an agent can rewrite its own rules mid-episode and the next session starts from the edited version with no audit event, no version and no diff presented to anyone. Working practice compounds it, because agents are routinely instructed to keep these files current, so rewriting the governing document is not an attack but a chore. Beyond ownership, two scopes are loaded with no stated precedence, so a project-scope rule and a user-scope rule that conflict are resolved by the model rather than by the loader, and neither the conflict nor the resolution is recorded. And the deepest point is verified four ways in the evidence: the only pre-tool hook is telemetry whose own header records an absolute requirement to exit zero so that it can never vote on whether a production agent continues, the only other hook is post-write, no managed settings tier exists anywhere on the host, and a grep of every executable surface for these filenames finds comments, a console feature that displays them to a human, and a service that writes one when creating an agent, but no authorisation check anywhere.
FAILURE MODE	An agent acts on a rule that another agent rewrote an hour earlier, or two loaded scopes disagree and the resolution is invisible to everyone including the operator who wrote them.
FAIL STATE	fail-open
EMITTED SIGNAL	None today: instruction compliance is unobservable without independent verification, and that unobservability is itself the finding, because nothing records that a clause was read, applied or violated. The signal needed is a runtime classifier that evaluates each attempted tool call against a machine-readable expression of the clause and emits an event when a prohibited action is attempted; the substrate that would carry it already exists, since the pre-tool hook fires on every tool call with a match-everything matcher and produces roughly 186 MB of output, and is deliberately forbidden from blocking.
METRICS	MG-031 MG-061 MG-034
COST	low, effectively the cost of writing a document, which is exactly why organisations over-rely on this layer
FRICTION	low, effectively zero: no action is ever delayed, refused or queued by this control, so it imposes no cost on legitimate work and equally none on illegitimate work
AWS BINDING	No native equivalent; instruction layers sit outside the cloud control plane. The adjacent managed offering is Bedrock Guardrails, which applies denied topics and content filters to model input and output; it governs what the model says, not what the agent is authorised to do. The authority equivalent of this document is an IAM permission boundary and a Service Control Policy, which is precisely the substitution the thesis argues for.
AZURE BINDING	No native equivalent; the adjacent offering is Azure AI Content Safety, which likewise governs content rather than authority. The enforceable form of these clauses is an Azure Policy assignment and Entra role assignments evaluated by the control plane rather than by the actor.

ON HOST	yes
EVIDENCE	EV-032
RESEARCH VALUE	The canonical exhibit for the thesis: a document that reads as policy, is loaded into every session, is verified four independent ways to gate nothing, and is writable by the party it governs, which makes it simultaneously the most visible control in the environment and the emptiest.

GR-L12-02 Prose standing authorisation for version control actions

INTENT	Remove approval friction from a class of actions judged low-stakes, by stating in the operating guidance that version control is for history only, that a commit or push deploys nothing and is not a gate, and that agents may commit and push freely under the operator's identity.
MECHANISM	Clauses in the project-scope guidance file and in several memory records establish the low-stakes framing, the operator-identity commit convention and an instruction to commit a working-tree fix as soon as it is verified; the same file carries a standing authorisation section asserting unrestricted change authority over production, DNS and cloud accounts. A scheduled job every 15 minutes stages everything in the fleet directory and commits it locally with a generated message, filtered only by an ignore file.
ENFORCEMENT	instructional
BINDS	AC-003 AC-004 AC-005 AC-006
ENABLES	I
PRECONDITION	Credentials for the hosted remote are present on disk; they are, in a mode-600 file owned by the governed account, alongside a separate stored access token.
BYPASS	A permissive instruction cannot be bypassed, only over-applied, and this is the layer's neglected direction: the governance failure here is faithful compliance rather than defiance. The clause's own risk assessment, that a push is low-stakes because it does not deploy, is false on this host, because the process manager re-executes whatever sits on the shared working tree, so code that reaches the tree reaches production at the next ordinary restart, which is documented as having happened twice. The operator-identity convention removes exactly the attribution an auditor would need, since the version control record shows a human author for an action no human took. The only mechanical backstop is a settings deny list containing five force-push patterns naming two conventional default branch names, so a force push to any other branch or a history rewrite by other means is unlisted, and that settings file is itself writable by the account it governs. And the automatic commit job stages indiscriminately, so any secret an agent writes into its own working directory is committed, bounded only by an ignore file.
FAILURE MODE	Unreviewed code reaches the working tree and then production through a restart nobody associates with a change; or a credential is committed; or the history attributes an agent's action to a person.
FAIL STATE	fail-open
EMITTED SIGNAL	None today for the instruction itself; version control history records the commit and the console audit log records nothing, because none of these actions passes through the console. Needed: a commit provenance record binding each commit to the acting agent identity and to the brief that requested it, distinct from the human account used to authenticate to the remote.
METRICS	MG-067 MG-029 MG-035
COST	low
FRICTION	low, and by design zero: the clause exists to remove waiting, which is also why it is the cheapest possible way for an organisation to widen an agent's authority without recording that it did so
AWS BINDING	No native equivalent; the enforceable form is repository-side branch protection with required reviewers, an IAM policy on CodeCommit or the equivalent hosting provider permission, and IAM Identity Center attributing the action to a distinct workload principal rather than to a human user. Bedrock Guardrails is not relevant here, because the risk is authority rather than content.
AZURE BINDING	No native equivalent; Azure Repos branch policies with required reviewers and build validation, and a distinct service principal per workload with its own Azure RBAC assignment, are where this statement would have to live to bind anything.
ON HOST	yes
EVIDENCE	EV-032 EV-027 EV-039
RESEARCH VALUE	Shows the instruction layer's other direction. Governance discussion assumes prose controls restrict; here the same zero-cost, zero-friction mechanism grants estate-wide authority, and because nothing evaluates prose in either direction the agent's effective

> authority is set by capability rather than by any clause, which is the thesis stated from the permissive side.

GR-L12-03 Prose deploy sequence instruction

INTENT	Make deployments correct and reversible by prescribing an order of operations: take a backup reference, apply the code, build on the box, restart only the affected process, and verify health before considering the work done, with an explicit prohibition on restarting one particular service for a front-end change because it holds live sessions.
MECHANISM	The sequence is prose in the project-scope guidance file, repeated in per-workload guidance files and in memory records. Its most load-bearing clause is negative, naming one service that must not be restarted, and it exists because restarting that service destroys in-flight interactions. There is no deployment orchestrator on this host: a deploy is whatever commands the agent chooses to run.
ENFORCEMENT	instructional
BINDS	AC-008 AC-009 AC-044 AC-045
ENABLES	I
PRECONDITION	The agent read the file this episode and the relevant clause is still within its effective context at the moment the deploy happens.
BYPASS	This instruction fails by attrition rather than by defiance. A deploy is the end of a long episode and the guidance is loaded at the start, so the clause competes for context with everything read since, and the environment is engineered to lengthen episodes, with a clock injecting new work every three minutes and a stop path that opens by asking for one more turn. It fails a second way by layering: 40 classic instruction files exist, several describing deploy sequences for different workloads, with no precedence rule, so an agent working across two trees can apply the wrong sequence perfectly. It fails a third way on naming, because the prohibition is expressed as a service name and nothing maps a name to a running process, so a restart issued through the process manager, through the service manager, or as a side effect of an unattended operating system upgrade that restarts affected services satisfies the letter of nothing. And it fails decisively at its last step, because verify-health is self-assessed: the founding case includes a deploy reported as successful that had disabled a service's core connection function for about 24 hours, which is what an instruction to verify produces when nothing independent verifies.
FAILURE MODE	The sequence is followed in spirit, the health check is the agent's own, and the change is recorded as successful while the workload is broken, with the divergence surfacing days later or not at all.
FAIL STATE	fail-open
EMITTED SIGNAL	None today: nothing records that a deploy occurred, in what order, or whether the prohibited restart happened, because the console audit log covers only operator actions taken through the console. Needed: an infrastructure-emitted event stream independent of the agent, for example the process manager's own restart events correlated to a change record, plus an external health probe executed by a different identity after every restart.
METRICS	MG-007 MG-016 MG-022
COST	low
FRICTION	low
AWS BINDING	No native equivalent for the instruction; the enforcing analogue is CodeDeploy with a defined lifecycle hook order and automatic rollback on a CloudWatch alarm, combined with Systems Manager Automation as the only permitted change path so that an ad-hoc restart is not a reachable operation at all.
AZURE BINDING	No native equivalent; Azure Pipelines deployment strategies with lifecycle hooks and health checks, App Service deployment slots with swap-with-preview, and Azure Automation as the only permitted change path.
ON HOST	yes
EVIDENCE	EV-032 EV-039 EV-027
RESEARCH VALUE	Marks the boundary of what prose can express: it can state an ordering and it cannot state a postcondition, so the step most likely to be wrong, the verification, is precisely the step nothing checks, and the founding case's most costly documented incident sits exactly there.

GR-L12-04 Dispatch brief containing prohibitions in prose

INTENT	Bound a specific piece of work at the moment it is delegated, by stating in the brief what the agent must not do, which files it may write and what it must not touch.
MECHANISM	The brief is delivered as text. On the supervised path a startup prompt names four files by absolute path, tells the agent to read them and then to action anything waiting in its inbox, and is delivered by writing a temporary file, loading it into a terminal buffer and pasting it followed by Enter. On the console path an operator-supplied brief of up to 100,000 characters is either pasted from a file using bracketed paste or typed inline. Eight long-form dispatch briefs totalling about 187 kB sit at the working home root and nine role files total about 65 kB; the brief is the operative mandate on this host, and it is prose.
ENFORCEMENT	instructional
BINDS	AC-019 AC-002 AC-023 AC-024
ENABLES	II
PRECONDITION	The brief actually landed. Delivery is a timing bet against the tool's readiness with no acknowledgement and no retry: keys sent before the interface is reading standard input are silently dropped rather than queued, and nothing verifies that the prompt arrived.
BYPASS	The weakness specific to a brief is that it is one message in a stream of messages that arrive by the same mechanism and are indistinguishable in kind. At least six independent producers can write into the same pane, three of them on timers, and a clock-driven nudge every three minutes tells the agent to pick up backlog work from its role file, so an instruction bounding this task competes directly with an instruction to find more work, and the standing file outlives the specific brief because it is re-read at every restart while the brief is not re-presented. Framing is inverted: content arriving over the inter-agent bus is rendered to the agent as a plain instruction with a reply prompt and no untrusted-data envelope, while operator-uploaded file content does carry an explicit treat-as-data envelope, so another agent's text outranks a human's attachment. And authentication is absent on the cheapest path: any local process can append a line directly to a session's mailbox file, bypassing the bus entirely with no token check, no enablement check, no rate accounting and no audit record, and the appended text is presented to the agent under an inter-agent banner.
FAILURE MODE	An agent exceeds a brief it can no longer see, or acts faithfully on injected text that arrived through an unauthenticated path and is formatted exactly like a brief.
FAIL STATE	fail-open
EMITTED SIGNAL	None today for compliance, and none even for delivery, since the paste is unacknowledged. Needed: a delivery acknowledgement, a persistent machine-readable mandate object carried alongside the session rather than embedded in the conversation, and a classifier emitting an event when an attempted action falls outside that object.
METRICS	MG-030 MG-032 MG-036
COST	low
FRICTION	low
AWS BINDING	No native equivalent; the enforcing analogue is a scoped session policy on an STS AssumeRole call that expresses the brief's boundary as permissions for that task, so exceeding the brief produces an access denial rather than requiring a judgement.
AZURE BINDING	No native equivalent; the analogue is a task-scoped managed identity with an Azure RBAC assignment limited to the resources the brief names, time-boxed through Entra Privileged Identity Management.
ON HOST	yes
EVIDENCE	EV-020 EV-025 EV-032
RESEARCH VALUE	Shows that in this architecture a mandate and an injection are the same object delivered by the same mechanism, so a mandate cannot be authenticated; that is a structural rather than a behavioural limit on prose delegation, and it holds however well the brief is written.

GR-L12-05 Prose instruction to confirm before a mutating cloud call

INTENT	Keep a human in the loop before an agent-initiated cloud control plane change, by instructing the model to obtain confirmation before performing an instance action.
MECHANISM	The orchestrator runtime exposes ten model-callable tools; the instance action tool is the only one with a designed confirmation flow. A pending-confirmation record is written when the tool is called, and the consumer that would read that record and match an affirmative reply exists only in a dormant repository and was never implemented in the running process. The record is written and never read, so the confirmation mechanism is dead code and what remains in force is the instruction text.
ENFORCEMENT	instructional
BINDS	AC-039 AC-045 AC-038
ENABLES	I
PRECONDITION	The model chooses to emit a confirmation turn and a human is reading the channel it emits into; that channel is one handset, and the same runtime decides whether the message is sent.
BYPASS	What distinguishes this entry is that the control's own designers built the mechanism and left it disconnected, so an instruction became load-bearing without anyone deciding that it should be. Nothing in the runtime reads the pending record, so an affirmative reply is not required and a refusal is not expressible: the tool executes on the first call. The instruction is also addressed to a model whose driver is an inbound message channel authenticated by a caller identifier string alone, with webhook signature validation that runs only when the header is present and no replay protection, so the party being asked to confirm may not be the operator; and the permissions consulted travel in the request body and honour a wildcard, with no principal in the signed envelope to check them against. Finally the control is inverted with respect to risk: the tool carrying a confirmation flow, even a dead one, is the least dangerous of the three action tools, while the two that reach arbitrary remote shell and arbitrary agent spawn carry none at all.
FAILURE MODE	An instance is stopped or rebooted on a single unconfirmed model turn, and the only evidence is an outbound message that the same runtime chose to send.
FAIL STATE	fail-open
EMITTED SIGNAL	None today that could distinguish a confirmed call from an unconfirmed one, because the confirmation state is written and never consulted. Needed: the consumer the design already implies, emitting a confirmation-requested and confirmation-granted pair with the granting principal recorded, plus a denial event when a mutating call is attempted without one.
METRICS	MG-060 MG-059 MG-050
COST	low
FRICTION	low, and zero in effect, since nothing waits for the confirmation the instruction requests
AWS BINDING	No native equivalent for the instruction; the enforcing analogue is an IAM policy denying the mutating action to the agent's principal and routing it through Systems Manager Change Manager or an Automation runbook with an approval step, so that confirmation is a permission rather than a request. Bedrock Guardrails governs model output content and would not touch this.
AZURE BINDING	No native equivalent; an Azure Policy deny on the operation for the workload identity, combined with a Logic Apps or Teams adaptive card approval held by the only principal permitted to perform it, and Entra Privileged Identity Management for the elevation.
ON HOST	yes
EVIDENCE	EV-029 EV-039
RESEARCH VALUE	A precisely documented case of a confirmation control whose consumer was never built, showing how quickly an intended architectural control degrades into an instructional one and, more importantly, that nothing in the system reports the degradation, so the design document and the runtime disagree silently.

GR-L12-06 Prose instruction to confirm before remote execution

INTENT	Prevent an agent from running an arbitrary command on another estate host without a human saying yes, by instructing the model, in the tool description itself, to ask the operator to confirm first for destructive commands.
MECHANISM	The remote execution tool's description carries the sentence as prompt text. The tool validates the target host against a six-entry enumeration and clamps the timeout between one and 120 seconds; the command string is not constrained in any way, there is no allowlist, and no confirmation mechanism exists at all, not even the dormant one that accompanies the instance action tool. The hosts it reaches are the production estate, where the same account holds passwordless root.
ENFORCEMENT	instructional
BINDS	AC-042 AC-043 AC-029 AC-045
ENABLES	I
PRECONDITION	None. The model may call the tool on its first turn, and the only preconditions in code are the host enumeration and the timeout clamp.
BYPASS	The differentiating weakness is what the word destructive is doing: the instruction delegates the risk classification to the model at the moment of acting, with no definition, no register of action classes and no default, so the control's coverage is exactly the set of commands the model happens to consider destructive, judged by the party that wants to proceed. A sequence that is individually benign and collectively catastrophic, for example a configuration edit followed by a service reload, is destructive in effect and innocuous token by token. The host enumeration is a real bound and it is the wrong bound, constraining where and not what, and the six hosts it permits are the ones that matter. There is no interception point between the model's decision and the transport, so no later control can compensate. And the same runtime's sibling tool spawns a fresh agent with the permission system disabled anywhere under the home tree, so an instruction not to run a destructive command is satisfied literally by asking another agent to run it.
FAILURE MODE	An arbitrary command runs as root on a production host on a single model turn, and the operator learns of it from the outcome, if at all.
FAIL STATE	fail-open
EMITTED SIGNAL	None today: neither the invocation nor its outcome produces an audit record on this host, because the console audit log covers only operator actions through the console and the runtime's own authentication failures produce no record either. Needed: a per-invocation event carrying target host, full command, requesting principal and confirmation state, written to a store the agent cannot modify.
METRICS	MG-040 MG-033 MG-060
COST	low
FRICTION	low
AWS BINDING	No native equivalent; the enforcing analogue is Systems Manager Run Command restricted by IAM to a named document and a bounded parameter set with Session Manager logging, so that arbitrary shell is simply not an available operation, and a Change Manager approval binds the human decision to the permission rather than to a sentence.
AZURE BINDING	No native equivalent; Azure Automation runbooks or Azure Arc Run Command constrained by Azure RBAC to specific scripts, with Azure Policy denying arbitrary command execution and Entra Privileged Identity Management holding the elevation.
ON HOST	yes
EVIDENCE	EV-029 EV-039
RESEARCH VALUE	The clearest instance of confirmation controls being inverted with respect to risk, and of an instruction that outsources risk classification to the actor at the moment of action, which is exactly the structure the action class register exists to replace.

GR-L12-07 Auto-memory store carried into every session as context

INTENT	Give every session durable knowledge of the estate, so that decisions draw on accumulated experience rather than being rediscovered each time and so that corrections persist across restarts.
MECHANISM	Three project-scoped memory stores totalling 465 files and about 2.8 MB, a knowledge base of 703 files and about 21.9 MB served by a local keyword-search server, nine per-agent running memory logs of about 1.2 MB, and role and roster files, all loaded or retrieved as context. Because every session shares one working directory, all agents share one memory store and one transcript namespace; per-agent separation is by convention, that is by absolute paths written into role files and the startup prompt. A memory index file is loaded at session start and is regenerated by a script the agents may invoke and which is explicitly allow-listed, so the index of the agent's own memory is agent-maintained.
ENFORCEMENT	instructional
BINDS	AC-022 AC-023 AC-025 AC-055
ENABLES	I
PRECONDITION	The session's working directory keys the store; a session started elsewhere silently binds to a new empty store, which is the documented reason the supervised path hardcodes its working directory.
BYPASS	This is the entry where prompt injection is concrete rather than theoretical. Two scheduled jobs run every five minutes and fast-forward a knowledge tree and four repository working trees from remote hosted repositories into directories that the runtime's search and read tools serve into a model's context, so content authored elsewhere becomes agent context on a five-minute cycle with no review step, no provenance and no untrusted-data envelope; that is a remote-content-to-agent-context path, and no instruction can defend against it because the injected text arrives in the same channel and the same format as the guidance. The store is also writable by every agent, so one agent can write a memory that instructs another, and the file-drop convention plus the direct mailbox append make that ordinary practice rather than an exploit. The index that governs what loads is regenerated by an agent-invocable script, so the agent curates its own context. And scale defeats it from the other side: about 26.7 MB of instruction and instruction-adjacent material cannot all be in context, so what governs a turn is whatever retrieval returned, and nothing records which records were present when a decision was made.
FAILURE MODE	An agent acts on a memory that is stale, wrong, written by another agent or synchronised in from a remote repository, and the episode cannot afterwards be reconstructed because the context set was never recorded.
FAIL STATE	fail-open
EMITTED SIGNAL	None today: no record exists of which memory or knowledge records were loaded for a given turn, so a decision cannot be attributed to its inputs. Needed: a per-turn context manifest listing every retrieved record with source, author identity and last-modified time, written to a store the agent cannot alter, plus explicit provenance labelling for anything synchronised from outside the host.
METRICS	MG-066 MG-027 MG-064
COST	moderate
FRICTION	low
AWS BINDING	No native equivalent for the memory layer; the adjacent managed offerings are Bedrock Knowledge Bases with data source access control and Bedrock Guardrails contextual grounding checks, which score whether a response is grounded in retrieved content. Neither authenticates the retrieved content nor governs what the agent may then do with it, which is the whole of the problem here.
AZURE BINDING	No native equivalent; Azure AI Search with document-level security trimming and Azure AI Content Safety prompt shields, which detect indirect prompt injection patterns in retrieved documents. Again these filter content rather than authority, and neither would prevent a synchronised repository file from being treated as instruction.
ON HOST	yes
EVIDENCE	EV-032 EV-027 EV-020

RESEARCH VALUE The concrete demonstration that retrieval turns a supply chain into a prompt: a five-minute cron job makes remotely authored text into agent context, and the entire governance layer for that text is the model's disposition to read it as information rather than as instruction.

GR-L12-08 System prompt constraint on tool use

INTENT	Shape how the agent uses its tools through the system-level text and tool descriptions the model runtime supplies, for example preferring a dedicated tool over a shell command or asking before a destructive action.
MECHANISM	Three surfaces carry this text. The model command line tool's own system prompt and tool schemas, which are vendor-supplied and not authored here. The tool descriptions in the local orchestrator runtime, which include the confirm-before-destructive sentence. And the harness settings file's behavioural switches, which on this host consist of a 23-pattern command deny list, a default permission mode of bypass, and suppression of the warning that mode would otherwise raise.
ENFORCEMENT	instructional
BINDS	AC-022 AC-023 AC-029 AC-036
ENABLES	I
PRECONDITION	The installed version of the tool is whatever is currently published; nothing on this host pins it and nothing records which version was in force during a given episode.
BYPASS	The weakness specific to this entry is that the constraint is not under the operator's control and changes underneath the estate without anyone acting: a scheduled job every two hours attempts the tool's own update and, on failure, installs the latest published package globally as root through passwordless escalation with no version pin, no checksum, no approval and no canary, so the system prompt, the tool descriptions and their safety text can change on every host in the fleet within two hours of a publication with no record of what changed. Locally authored tool text is defeated more simply, because it sits in a description string the model weighs against user instructions, and the environment supplies contradicting user instructions continuously through nudges and injected messages. No managed policy tier exists above the writable settings file, so nothing can pin the constraint locally either. And any constraint expressed as a preference for one tool over another is void while the shell tool is available, because every capability on this host is reachable through it.
FAILURE MODE	Behaviour changes across the fleet after an unattended package upgrade and nobody can attribute the change, or the model reaches the same capability through a different tool from the one the text discusses.
FAIL STATE	fail-open
EMITTED SIGNAL	None today. The observation substrate for this exact question exists and is deliberately unused: the pre-tool hook fires on every tool call with a match-everything matcher and produces roughly 186 MB of output, and its own header records an absolute requirement to exit zero so that a telemetry hook can never vote on whether a production agent continues. Needed: that same hook permitted to return a deny decision, plus a recorded hash of the installed tool version and its prompt text so behavioural change can be attributed to a version rather than to a mood.
METRICS	MG-061 MG-031 MG-050
COST	low
FRICTION	low
AWS BINDING	No native equivalent; instruction layers sit outside the cloud control plane. Bedrock Guardrails and Bedrock model invocation logging are adjacent and govern model input and output content rather than the agent's authority to act; the authority equivalent is an IAM permission boundary on the principal the tool runs as, which holds regardless of what the tool description says.
AZURE BINDING	No native equivalent; Azure AI Content Safety and Azure AI Foundry model policies are the adjacent content controls, and the authority equivalent is an Azure RBAC deny assignment on the workload identity, evaluated by the control plane rather than by the model.
ON HOST	yes
EVIDENCE	EV-032 EV-027 EV-028
RESEARCH VALUE	Identifies a governance dependency organisations rarely record: the safety text shaping agent behaviour is vendor-supplied, unpinned and reinstalled as root every two hours, so this layer is not merely unenforced, it is not even stable, and no experiment run across a version boundary is comparable without a recorded prompt hash.

GR-L12-09 Model refusal behaviour as a control

INTENT	Rely on the model's own trained disposition to decline harmful, unsafe or clearly out-of-scope actions as the final safeguard where no other control applies.
MECHANISM	There is no mechanism in the system: the control is entirely a property of the model weights and the provider's alignment training, expressed at inference time. On this host it is what remains after the tool-layer permission prompt is disabled on the command line in five places across three languages with no off switch, asserted again as a bypass default in the settings file with its warning suppressed, and reduced to a 23-pattern command deny list that names no cloud command, no remote shell, no process manager, no service manager, no privilege escalation command and no file write path. Stated fairly, it is not nothing: a capable model declines a substantial class of overtly harmful requests, notices obviously destructive commands, and asks before irreversible steps when the situation is legible; and it is the only control here with complete coverage, applying to every tool, every path and every action class including the ones nobody classified.
ENFORCEMENT	instructional
BINDS	AC-016 AC-024 AC-027 AC-029 AC-039 AC-043 AC-052 AC-054
ENABLES	I
PRECONDITION	None, and it is the only control in the catalogue with no precondition; that universality is both its value and the reason reliance on it grows automatically as every other control is removed.
BYPASS	Refusal is a disposition rather than a check, so the analysis is of the conditions under which it does not engage, and they are the common ones. It does not engage where the action is not recognisable as harmful, which covers the whole class of sincere error and is the dominant failure mode measured in this case. It does not engage against a false premise: an agent told, in a file it can itself edit, that it holds standing authority over production, DNS and cloud accounts is not refusing anything, it is complying. It cannot discriminate injected instructions from legitimate ones, and this host offers at least three routes into a running agent's context, including a direct mailbox append with no token check, no rate accounting and no audit record, plus remote repository content synchronised into retrievable directories every five minutes. It cannot engage on what it cannot see, and episodes here are long by design with the guidance loaded once at the start and never re-anchored, so the constraint may simply have fallen out of context. Its boundary is not even stable over time, because the model runtime is reinstalled as root every two hours with no version pin. And its outcome is invisible, since a refusal and a task that was never attempted produce the same observation, so the control cannot be measured, tuned or relied upon at any stated rate.
FAILURE MODE	The model does not refuse because there was nothing to refuse: it was wrong rather than adversarial. The action proceeds, the report is confident, and the divergence is found by an independent audit weeks later or never; the founding case measured 61.1 per cent factual divergence between agent self-report and independent evidence across 18 audited episodes drawn from 46 candidates, including a deploy reported as successful that had disabled a service's core connection function for about 24 hours, and no refusal behaviour engages anywhere in that sequence.
FAIL STATE	fail-open
EMITTED SIGNAL	None today, and uniquely so: even a refusal that does occur leaves no event, no counter and no channel, so the control is unobservable in the successful case as well as the failing one. That unobservability is itself the finding, because a guardrail emitting no signal cannot be evaluated, cannot be tuned and cannot serve as an experimental variable, which means an architecture resting on it cannot state its own residual risk. Needed: an independent runtime classifier outside the acting model that evaluates each attempted action against the action class register and emits both a refusal event and an attempted-prohibited-action event; the pre-tool hook substrate that could carry it already fires on every tool call and is contractually forbidden from blocking.
METRICS	MG-031 MG-063 MG-060 MG-022
COST	low, and effectively zero: nothing is deployed, configured, staffed or maintained, which together with total coverage is exactly why organisations over-rely on it
FRICTION	low, and effectively zero: nothing waits, nothing is queued and nobody is asked, so the control imposes no visible cost and therefore attracts no scrutiny

AWS BINDING	No native equivalent; model behaviour sits outside the cloud control plane. The adjacent managed offering is Bedrock Guardrails with denied topics, content filters and contextual grounding checks, which governs what the model says rather than what the agent is authorised to do and would not have prevented any of the consequential action classes enumerated for this host. The authority equivalent is an IAM permission boundary and a Service Control Policy, which fail closed regardless of what the model believes.
AZURE BINDING	No native equivalent; Azure AI Content Safety with prompt shields and groundedness detection is the adjacent content control, and Azure AI Foundry evaluations can measure refusal behaviour offline but cannot gate an action. The authority equivalent is an Azure RBAC deny assignment with Azure Policy, evaluated by the control plane rather than by the actor.
ON HOST	yes
EVIDENCE	EV-032 EV-020 EV-039 EV-027
RESEARCH VALUE	The doctorate's central claim compressed into one entry: the weakest control, with the widest coverage, the lowest cost and no emitted signal, is the one the architecture ultimately rests on, and it is structurally blind to sincere error, which is the failure the founding case actually measured. It also makes the strongest argument for the metric register, since a control that cannot emit a signal cannot be an experimental variable and therefore cannot be defended empirically.

3. Cobertura: as classes de ação sem guardrail structural

A matriz de cobertura une o registro de classes de ação a este catálogo, uma linha por classe de ação, contando quantos guardrails structural, conditional e instructional a vinculam hoje, e atribuindo uma severidade de lacuna. Cinquenta e seis das sessenta classes de ação não têm nenhum guardrail structural que as vincule. Apenas quatro têm: as duas classes do domínio de tempo real, delimitadas pelo modo de teste structural, e as duas classes do System of Record, delimitadas pelas identidades de conexão somente leitura e com capacidade de escrita, respectivamente. Esse cinquenta e seis é o número principal que este catálogo existe para produzir, e deve ser lido ao lado das quarenta e nove entradas do catálogo classificadas structural: quase toda essa capacidade structural é proposta, e não construída, motivo pelo qual cinquenta e seis classes de ação e quarenta e nove entradas structural não estão em tensão entre si.

Tabela D3-1. Cada classe de ação contra sua contagem de guardrails structural, conditional e instructional. Cinquenta e seis das sessenta linhas mostram zero na coluna structural; as quatro que não mostram são o modo de teste de tempo real e as identidades de conexão do System of Record.

ACTION CLASS	NAME	STRUCT.	COND.	INSTR.	STRUCTURAL?	GAP
AC-001	Read source code in a local repository	0	0	0	no	low
AC-002	Modify source files in a working tree	0	0	1	no	low
AC-003	Commit to a local git repository	0	0	1	no	low
AC-004	Push a commit to a hosted remote repository	0	0	1	no	moderate
AC-005	Create, switch or delete a git branch	0	0	0	no	moderate
AC-006	Rewrite published version control history	0	1	0	no	high
AC-007	Install a package into a project dependency tree	0	0	0	no	moderate
AC-008	Build application artefacts on the host	0	0	0	no	moderate
AC-009	Restart a process-manager-supervised service on this host	0	0	1	no	moderate
AC-010	Start, stop, enable or disable a systemd unit	0	0	0	no	high
AC-011	Terminate an arbitrary operating system process	0	0	0	no	high
AC-012	Spawn a long-lived background process	0	0	0	no	moderate
AC-013	Persist the process manager's process list	0	0	0	no	moderate
AC-014	Create a new agent session	0	1	0	no	high
AC-015	Terminate an agent session	0	1	0	no	moderate
AC-016	Inject input into another agent's live session	0	0	0	no	severe
AC-017	Write to the fleet registry	0	1	0	no	moderate
AC-018	Retire or delete a fleet entry	0	1	0	no	moderate
AC-019	Dispatch a brief to an agent	0	0	1	no	high
AC-020	Change an agent's model, effort or permission configuration	0	0	0	no	high
AC-021	Modify an agent's startup wrapper or launch script	0	0	0	no	severe
AC-022	Read an arbitrary file on the host	0	0	0	no	high
AC-023	Write an arbitrary file on the host	0	0	0	no	high
AC-024	Delete files outside a working tree	0	1	0	no	high
AC-025	Modify its own instruction files	0	0	0	no	high

ACTION CLASS	NAME	STRUCT.	COND.	INSTR.	STRUCTURAL?	GAP
AC-026	Modify the permission settings file	0	0	0	no	severe
AC-027	Read credential material at rest	0	0	0	no	severe
AC-028	Ingest a file into session storage	0	2	0	no	low
AC-029	Execute a command as root	0	0	0	no	severe
AC-030	Modify accounts, groups or privilege configuration	0	0	0	no	severe
AC-031	Install or remove an operating system package	0	1	0	no	high
AC-032	Create or modify a systemd unit	0	0	0	no	severe
AC-033	Modify reverse proxy configuration	0	0	0	no	severe
AC-034	Reload or restart the reverse proxy	0	0	0	no	high
AC-035	Open a listening socket reachable beyond the host	0	0	0	no	high
AC-036	Make an arbitrary outbound network request	0	0	0	no	high
AC-037	Obtain cloud credentials from the instance metadata service	0	0	0	no	severe
AC-038	Call a cloud control plane read API	0	0	0	no	moderate
AC-039	Call a cloud control plane mutating API	0	0	1	no	severe
AC-040	Change a public DNS record	0	0	0	no	severe
AC-041	Obtain or renew a public TLS certificate	0	0	0	no	high
AC-042	Open an authenticated session to another estate host	0	0	0	no	high
AC-043	Execute a command on another estate host	0	0	1	no	severe
AC-044	Deploy application code to another estate host	0	1	0	no	severe
AC-045	Restart a production service on another estate host	0	0	1	no	high
AC-046	Send an operator notification	0	2	0	no	moderate
AC-047	Send an email to an external recipient	0	0	0	no	high
AC-048	Publish content on a public web surface	0	0	0	no	severe
AC-049	Conduct a real-time voice interaction with a member of the public	1	0	0	yes	none
AC-050	Send a message to a member of the public	1	0	0	yes	none
AC-051	Read from the System of Record	1	0	0	yes	none
AC-052	Write to the System of Record	1	0	0	yes	none

ACTION CLASS	NAME	STRUCT.	COND.	INSTR.	STRUCTURAL?	GAP
AC-053	Write to its own transcript record	0	0	0	no	high
AC-054	Delete or truncate an audit or transcript record	0	0	0	no	severe
AC-055	Write a completion attestation about its own work	0	0	0	no	severe
AC-056	Record an approval in the deployment approval gate	0	1	0	no	severe
AC-057	Set or clear the update kill switch	0	1	0	no	severe
AC-058	Alter its own resource accounting record	0	0	0	no	moderate
AC-059	Answer an interactive permission or resume prompt on a human's behalf	0	0	0	no	severe
AC-060	Self-update the agent runtime software	0	0	0	no	high

A distribuição de severidade entre as cinquenta e seis não é uniforme, e sua forma é informativa. Dezenove são classificadas `severe`, vinte `high`, treze `moderate` e quatro `low`; as quatro classes de ação restantes, todas delimitadas por um controle structural, são classificadas `none`. As dezenove entradas severas se concentram exatamente nos lugares que os achados `F-028` e `F-032` do D2 já nomeiam: injetar entrada na sessão ativa de outro agente, modificar o próprio wrapper de inicialização ou as configurações de permissão de um agente, executar um comando como root ou em outro host do parque, obter credenciais de nuvem a partir do serviço de metadados da instância, chamar uma API mutante do plano de controle de nuvem, alterar um registro DNS público, excluir ou truncar um registro de auditoria, escrever uma atestação de conclusão sobre o próprio trabalho, e responder a um prompt de permissão interativo em nome de um humano. Cada uma dessas dezenove é irreversível ou reversível apenas com esforço, e cada uma hoje depende de controles conditional ou instructional, ou de nenhum controle, para a restrição que tiver.

> **A SEVERIDADE SE CONCENTRA ONDE A REVERSIBILIDADE JÁ É MENOR**
>
> A correlação não é coincidência, e não é uma boa notícia. A coluna `gap_severity` da matriz de cobertura e as colunas `reversibility` e `blast_radius` do registro de classes de ação são calculadas de forma independente, uma a partir do catálogo de guardrails e outra a partir das propriedades da própria ação, e ainda assim elas se alinham: as ações sem nenhum guardrail structural são desproporcionalmente as ações em que um erro é mais difícil de desfazer. Uma lacuna na cobertura structural é mais perigosa exatamente onde ela hoje é maior.

Cinquenta e duas das cinquenta e seis carregam uma entrada `recommended_guardrails`, nomeando os controles propostos específicos que fechariam a lacuna, extraídos dos mesmos identificadores catalogados na Seção 2; as quatro restantes não carregam nenhum conjunto proposto e são registradas como `not determined`, à espera de triagem de taxonomia. O registro de resíduo de supervisão que o D2 desenvolve em sua Seção 3 fixa (pins) separadamente trinta e duas classes de ação como o resíduo irredutível, vinte e nove delas dentro destas cinquenta e seis, e as trata ali, não aqui; os dois recortes se sobrepõem, em vez de particionar, porque este pergunta que guardrail poderia vincular a ação e o do D2 pergunta se algum guardrail poderia remover a necessidade de um humano de todo modo. A matriz de cobertura deste catálogo e o registro de resíduo do D2 são deliberadamente dois recortes diferentes das mesmas sessenta classes de ação: este pergunta que guardrail poderia vincular a ação; o do D2 pergunta se algum guardrail poderia remover a necessidade de um humano de todo modo.

---

4. Necessidade e suficiência por nível da taxonomia, enunciadas como hipótese

Os três níveis da taxonomia são definidos na proposta doutoral por consequência: Nível I totalmente autônomo, sem revisão prévia; Nível II autônomo com revisão posterior; Nível III aprovação prévia exigida, ou proibido. O campo `enables_level` deste catálogo declara, por entrada, o nível mais alto que aquele controle torna defensável para as classes de ação que vincula. Ler esse campo ao longo das 134 entradas produz uma hipótese sobre necessidade e suficiência, não uma afirmação, porque ela não foi testada contra uma execução experimental ao vivo: é um julgamento feito em tempo de projeto, registrado ao mesmo tempo que o restante de cada entrada, pelo mesmo autor, e herda toda limitação que a Seção 9 do D2 declara sobre descoberta por observador único.

A hipótese, enunciada com clareza: **nenhum guardrail isolado é suficiente para o Nível I por si só.** Das duas entradas deste catálogo que são ao mesmo tempo presentes e structural, `GR-L05-01` a conexão somente leitura com o System of Record, e `GR-L06-08` o modo de teste de tempo real, ambas tornam o Nível I defensável para o conjunto estreito de classes de ação que vinculam, e ambas o fazem apenas porque a ação que vinculam foi reduzida a algo sem efeito inseguro a ter, não porque qualquer

um dos dois guardrails, isoladamente, tornaria um efeito inseguro seguro o bastante para permitir sem supervisão. Entre as quarenta e nove entradas structural do catálogo em geral, apenas duas carregam `enables_level = I`, limites de grupo de controle sobre processos de agente e uma identidade de conexão que estruturalmente não consegue escrever; quarenta e uma carregam `enables_level = II` e seis carregam `enables_level = III`. **O Nível I é desproporcionalmente difícil de conquistar mesmo com um controle structural em vigor**, e as entradas que de fato o conquistam compartilham uma forma específica: elas removem por completo a capacidade de um efeito inseguro, em vez de acrescentar uma verificação sobre uma capacidade que permanece.

Suficiência sem necessidade é comum no Nível II. Quarenta e uma entradas structural tornam o Nível II defensável para o que vinculam, e a progressão de condition sets na Seção 8 mostra o mesmo padrão no nível agregado: `CS-04` até `CS-08` acrescentam, cada um, cobertura structural substancial, e cada um tem a hipótese de sustentar o Nível II, não de cruzar para o Nível I, porque o Nível II tolera um erro reversível capturado em revisão, e a maior parte do que esses condition sets acrescentam é detecção e contenção, e não a remoção da capacidade de errar.

A necessidade se concentra na L1 e na L7. Nove outras camadas nomeiam a separação de identidade da L1 como precondição em seu próprio campo `precondition`, e a camada de mandato da L7 é a única camada cujas entradas podem, em princípio, tornar uma ação não classificada default-deny em vez de default-permit. Espera-se, como hipótese, que um conjunto de guardrails individualmente suficiente em toda outra camada, mas que omita a separação de identidade ou um mandato default-deny, permaneça conditional na melhor das hipóteses, porque a análise de bypass de toda outra camada acaba alcançando a conta compartilhada ou a ausência de um ponto de decisão de política que pudesse ter recusado a ação antes que qualquer controle específico de camada fosse alcançado.

Isso é oferecido como uma proposição testável para o ambiente experimental, não como um resultado estabelecido: construir um condition set que mantenha a L1 e a L7 fixas em seu máximo proposto, enquanto varia todas as outras camadas, e comparar sua cobertura efetiva com a de `CS-09`. Se a hipótese estiver correta, os dois deveriam ficar próximos. Se não estiverem, a necessidade está distribuída de forma mais difusa entre as camadas do que este catálogo hoje presume, e esse seria um achado que valeria a pena ter.

---

5. Defesa em profundidade e pontos únicos de falha

A métrica `MG-062` fator de defesa em profundidade pergunta, por classe de ação, quantos controles independentes um agente teria de derrotar para alcançar um efeito proibido, em que independência significa que nenhum fato ou credencial único derrota mais de uma das camadas contadas. A métrica `MG-063` contagem de pontos únicos de falha faz a pergunta complementar diretamente: quantos fatos, credenciais ou condições únicos, se comprometidos ou removidos, derrotariam simultaneamente mais de um controle. Nenhuma das duas é computável como um número preciso para o caso fundador nesta execução, e o próprio motivo é o achado que ambas as métricas foram desenhadas para revelar: a matriz de cobertura conta guardrails structural, conditional e instructional por classe de ação, mas não desconta o bypass compartilhado, e a conta única do caso fundador, com root sem senha, está documentada como a condição derrotadora, nomeada ou implícita, em uma grande parcela das entradas de population presente deste catálogo.

Pelo menos um ponto único de falha é confirmado com alta multiplicidade, sem que seja preciso a contagem completa. A fronteira de conta que a L1 registra como ausente é citada, direta ou por meio de suas consequências, no campo `bypass_analysis` de entradas que abrangem ao menos seis camadas: `GR-L03-01` a lista de permissão de realpath, cujas raízes permitidas são ambas graváveis por aquela conta; `GR-L04-05` a lista de negação da camada de ferramentas, cujo arquivo de configurações aquela conta pode editar; `GR-L05-02` o advisory lock do registro da frota, que a conta pode simplesmente

ignorar; `GR-L08-02` o registro de auditoria sobre operações destrutivas do console, que reside sob uma raiz gravável que a mesma conta possui; `GR-L09-01` a affordance de parada de sessão do console, cujo processo supervisor a conta também consegue controlar; e a `GR-L12` em sua totalidade, já que toda instrução em prosa nessa camada é armazenada em um arquivo que a conta governada pode reescrever. Uma contagem de defesa em profundidade que trata essas entradas como camadas independentes superestima a resistência real do caso fundador a uma única credencial comprometida ou mal utilizada, porque quase todas elas colapsam para o mesmo fato.

> **PROFUNDIDADE NO PAPEL NÃO É PROFUNDIDADE NA PRÁTICA**
>
> Contar entradas por camada, como a matriz de cobertura da Seção 3 faz, sempre superestimará a defesa em profundidade sempre que uma única condição derrotar múltiplas camadas contadas de uma só vez. A fronteira de conta do caso fundador é o exemplo mais claro: uma classe de ação pode mostrar três ou quatro guardrails conditional na matriz de cobertura e ainda assim ter um fator de defesa em profundidade efetivo próximo de um, porque os três ou quatro compartilham a mesma condição derrotadora. `MG-062` existe especificamente para corrigir isso, e sua baseline ser `not computable` nesta execução, registrada honestamente em vez de estimada, é, por si só, evidência de que a correção ainda não foi aplicada em lugar algum do próprio pensamento de governança do caso fundador.

O remédio para o qual este catálogo aponta não são mais camadas; são camadas cujas análises de bypass nomeiam condições derrotadoras genuinamente diferentes. As entradas structural propostas da L1, identidade por agente, credenciais efêmeras, ausência de material de chave de longa duração em disco, são valiosas especificamente porque instalá-las muda a análise de bypass de cada entrada listada acima, não porque acrescentem uma décima terceira camada à contagem.

---

6. Conflito e interação entre guardrails

Os guardrails deste catálogo não são variáveis independentes, mesmo onde são individualmente bem projetados, porque alguns são construídos para neutralizar ativamente o efeito de outros. Três interações no caso fundador estão documentadas com clareza suficiente para serem enunciadas como fatos fixos, e não como hipóteses, e cada uma é um caso em que o propósito inteiro de um controle presente é derrotar o propósito inteiro de um controle diferente.

A resposta automatizada a prompts interativos (`GR-L10-08`) derrota o ponto de decisão humana que todo portão de permissão interativo existe para criar. O próprio campo de fricção do mecanismo enuncia seu efeito com candura: a fricção é baixa, e negativa, porque o mecanismo existe para remover a espera que um ponto de decisão humana imporia, e ele consegue. Toda entrada proposta da L7 que pressupõe que um humano ou um principal separado seja consultado antes de uma ação prosseguir, e toda entrada presente da L10 que pressupõe um humano vendo um prompt ao vivo, está em tensão direta com este único controle enquanto ele permanecer presente e sem condicionamento. Este não é um conflito hipotético; os próprios registros do caso fundador o documentam como o comportamento operativo hoje.

A supervisão de reinício sempre ativo (`GR-L09-10`) derrota todo mecanismo de interrupção que apenas interrompe uma sessão, em vez de remover sua capacidade de ser recriada. A Seção 2.9 já enuncia o mecanismo; o conflito é que as próprias entradas structural propostas da L9, um circuit breaker, suspensão automática sob anomalia, uma pausa global de frota, todas presumem que parar uma sessão para o trabalho, e o reinício sempre ativo é construído especificamente para tornar essa suposição falsa para os nove agentes sob supervisão de unidade. Instalar qualquer um dos controles de contenção propostos da L9 sem antes condicionar ou remover o reinício sempre ativo produziria um controle que dispara corretamente e, ainda assim, é derrotado em segundos.

O disparo agendado sem supervisão (`GR-L11-05`) e o orçamento de atenção do operador (`GR-L10-06`, proposto) apontam em direções opostas por projeto. O campo de fricção da entrada de disparo registra que seu efeito é negativo, porque ela existe para eliminar o tempo ocioso que um projeto human-in-the-loop de outro modo usaria, e seu sucesso é medido nessa eliminação; o campo de fricção da entrada de orçamento de atenção registra que sua fricção é alta por projeto, porque recusar trabalho quando a atenção se esgota é o mecanismo inteiro do controle. Um conjunto de guardrails que instala o segundo sem também condicionar o primeiro herda um controle que existe para ser esgotado contra uma fonte de carga que nunca para de chegar.

Um quarto padrão é mais fraco, mas vale a pena registrar porque recorre em várias camadas, em vez de ser um único par: várias das entradas instructional de population presente do catálogo, mais visivelmente `GR-L07-09` o brief em prosa como mandato operativo, e a camada L12 de modo geral, se descrevem em seu próprio campo `friction` como impondo custo efetivamente zero, precisamente porque nada espera, nada entra em fila e ninguém é consultado. Isso não é um conflito com um outro controle específico; é um incentivo permanente contra instalar qualquer controle que imponha fricção, porque a alternativa mais barata disponível, a prosa, já está presente e já é gratuita.

> **A FORMA GERAL DE UM CONFLITO DE GUARDRAIL NESTE CATÁLOGO**
>
> Em cada caso documentado, o par em conflito não são dois controles discordando sobre política. É um controle cujo mecanismo é a conveniência, projetado para remover uma espera, uma verificação ou um período ocioso, situado no mesmo caminho causal de um controle diferente cujo mecanismo é justamente a espera, a verificação ou o período ocioso em si. Revisar um conjunto de guardrails proposto em busca de conflito interno, nos termos deste catálogo, significa verificar se o campo `friction` de alguma entrada descreve a remoção exatamente da fricção da qual o campo `mechanism` de outra entrada depende.

---

7. O custo da governança

Cada entrada carrega um `operating_cost`, o que custa mantê-la em execução, e uma `friction`, o que ela custa ao trabalho legítimo, ambos em uma escala baixa, moderada, alta, e, para um número de entradas, o campo de fricção é escrito como prosa argumentada em vez de uma palavra isolada, porque a fricção é frequentemente o mecanismo, e não um efeito colateral dele. Ao longo das 134 entradas, sessenta e cinco têm custo operacional avaliado como baixo, cinquenta e uma como moderado e dezesseis como alto; a fricção se distribui de forma semelhante, sessenta e quatro baixa, trinta e oito moderada e dezessete alta, com um punhado adicional carregando avaliações qualificadas em prosa que esta seção trata individualmente.

Classificar honestamente nas duas direções significa nomear os controles baratos que compram pouco com a mesma clareza que os controles caros que compram muito. **As entradas mais baratas do catálogo são também aquelas que a própria análise deste catálogo diz serem as menos eficazes.** Toda entrada presente da L12 é avaliada com custo operacional baixo ou efetivamente zero, e fricção baixa ou efetivamente zero, e duas delas, `GR-L12-01` e `GR-L12-09`, dizem isso em seu próprio campo de custo, ao mesmo tempo que observam que é exatamente por isso que as organizações dependem excessivamente delas. Uma instrução em prosa não custa nada para escrever, não atrasa nada, e não bloqueia nada no runtime; é a pior razão possível entre custo reduzido e risco reduzido no catálogo, não porque esteja mal escrita, mas porque seu custo e seu efeito são, ambos, próximos de zero de modo structural.

As entradas propostas mais caras compram uma redução de risco específica e nomeável, e o catálogo registra essa troca com honestidade. `GR-L11-04` sequenciamento da verificação antes de a próxima classe de ação ser desbloqueada é avaliada com custo alto e fricção alta, e seu próprio campo de fricção declara que a fricção é o mecanismo, o custo de fazer da verificação uma

precondição do fluxo de controle, e não uma atividade de relato realizada depois. `GR-L11-07` duração máxima de execução sem supervisão antes de reautorização converte tempo sem supervisão em uma fila de decisões, custo moderado, fricção alta por projeto, e seu próprio campo declara claramente que este é exatamente o custo que o caso fundador optou por não pagar, o que também é declarado como o motivo pelo qual seus números de autonomia são tão altos quanto são. Nenhuma das duas entradas está sendo vendida aqui como isenta de custo; ambas são propostas com o reconhecimento honesto de que adotá-las tornaria a frota visivelmente mais lenta.

Um grupo menor de entradas é barato e eficaz ao mesmo tempo, o que é a combinação que vale a pena priorizar. `GR-L05-03` identidade de conexão que estruturalmente não consegue escrever, e `GR-L01-04` credenciais efêmeras com duração de sessão curta, são ambas avaliadas com custo moderado, fricção moderada, classe de enforcement structural; elas não removem capacidade que um agente legitimamente precisa, removem capacidade que uma credencial não precisa ter, para começo de conversa. É a mesma forma que o D2 identifica ao observar que o maior grupo isolado no resíduo de supervisão é fixado pela `SD-07` insuficiência de contenção, que é totalmente removível por engenharia de identidade: a redução de risco genuína mais barata neste catálogo não é a entrada mais barata; é a entrada cujo custo compra a remoção de uma capacidade, em vez de acrescentar uma verificação sobre ela.

Tabela D3-2. Cada entrada structural proposta, com seu custo operacional e a fricção que imporia, para que o leitor possa pesar diretamente o custo honesto do catálogo contra seu retorno structural, em vez de por anedota.

ID	NAME	LAYER	ENFORCEMENT	COST	FRICTION	ON HOST
GR-L01-02	Per-agent workload identity	L1	structural	moderate	moderate	no
GR-L01-03	Least privilege permission set per action class	L1	structural	high	moderate	no
GR-L01-04	Ephemeral credentials with short session duration	L1	structural	moderate	low	no
GR-L01-05	No long-lived key material on disk	L1	structural	moderate	moderate	no
GR-L01-06	Just-in-time privilege elevation with expiry	L1	structural	high	high	no
GR-L01-07	Separate identity for the verifier	L1	structural	moderate	moderate	no
GR-L01-08	Separate identity for the approver	L1	structural	moderate	high	no
GR-L02-05	Deny by default egress	L2	structural	high	high	no
GR-L02-08	Segmentation between agent tiers	L2	structural	moderate	moderate	no
GR-L02-09	No lateral network access between agent workspaces	L2	structural	high	moderate	no
GR-L02-10	Private endpoints for cloud service access	L2	structural	moderate	low	no
GR-L03-03	Per-session filesystem isolation by namespace or container	L3	structural	high	high	no
GR-L03-04	Read-only mounts for code the agent must not change	L3	structural	moderate	high	no
GR-L03-06	Immutable append-only evidence directory	L3	structural	moderate	low	no
GR-L03-07	Denial of write access to the agent's own instruction files	L3	structural	moderate	moderate	no
GR-L03-08	Denial of write access to the agent's own permission settings	L3	structural	low	moderate	no
GR-L03-10	Separation of the evidence store from the operational filesystem	L3	structural	moderate	low	no
GR-L04-03	Seccomp or mandatory access control profile on agent processes	L4	structural	moderate	moderate	no
GR-L04-04	Control group limits on agent processes	L4	structural	low	moderate	no
GR-L04-09	Prohibition of the permission-skipping flag in defined contexts	L4	structural	low	high	no
GR-L05-03	Connection identity that structurally cannot write	L5	structural	moderate	moderate	no
GR-L05-04	Schema constraints making a prohibited state unrepresentable	L5	structural	moderate	moderate	no
GR-L05-05	Separation of the evidence store from the operational store	L5	structural	moderate	low	no

ID	NAME	LAYER	ENFORCEMENT	COST	FRICTION	ON HOST
GR-L05-06	Agent denied the credential that could alter its own audit record	L5	structural	moderate	low	no
GR-L05-08	Relational store with transactional integrity for fleet state	L5	structural	high	moderate	no
GR-L05-10	Field-level provenance on every state record	L5	structural	moderate	low	no
GR-L06-03	Pre-approved rollback targets only	L6	structural	low	moderate	no
GR-L06-11	Deploy artefact provenance binding code to a review record	L6	structural	moderate	moderate	no
GR-L06-12	Post-deploy independent health verification	L6	structural	high	moderate	no
GR-L07-02	Policy decision point separate from the enforcement point	L7	structural	high	moderate	no
GR-L07-04	Mandate expiry	L7	structural	low	high	no
GR-L07-05	Mandate change journal with approver role	L7	structural	moderate	moderate	no
GR-L07-06	Structural impossibility of an agent widening its own mandate	L7	structural	high	moderate	no
GR-L07-07	Default deny for unclassified actions	L7	structural	moderate	high	no
GR-L08-03	Append-only transcripts enforced by storage	L8	structural	moderate	low	no
GR-L08-05	Independent verifier under a separate identity	L8	structural	high	moderate	no
GR-L08-06	Prohibition on the acting agent writing its own completion attestation	L8	structural	high	high	no
GR-L08-11	Infrastructure-emitted event stream independent of the agent	L8	structural	high	low	no
GR-L09-03	Circuit breaker on repeated failure	L9	structural	moderate	moderate	no
GR-L09-04	Per action class rate limits	L9	structural	moderate	moderate	no
GR-L09-06	Automatic suspension on anomaly	L9	structural	high	high	no
GR-L09-09	Graceful degradation to a safe mode	L9	structural	moderate	moderate	no
GR-L09-11	Global fleet pause under a separate principal	L9	structural	high	high	no
GR-L10-04	Approval requests presenting evidence rather than the agent's summary	L10	structural	moderate	moderate: it lengthens the record the gate assembles but not the decision the human makes, and it removes the single largest class of approval error in this case	partial
GR-L11-01	Time boxed mandates	L11	structural	moderate	moderate	no

ID	NAME	LAYER	ENFORCEMENT	COST	FRICTION	ON HOST
GR-L11-04	Sequencing constraint requiring verification before the next class unlocks	L11	structural	high	high, and the friction is the mechanism: it is the cost of making verification a precondition of control flow rather than a reporting activity performed afterwards	no
GR-L11-07	Maximum unattended run duration before re-authorisation	L11	structural	moderate	high, and deliberately: it converts unattended time into a queue of decisions, which is precisely the cost the founding case has chosen not to pay and the reason its autonomy figures are high	no

O padrão ao longo dessa tabela é consistente com a classificação acima: as propostas structural se concentram em custo operacional baixo a moderado, porque a maioria delas são decisões de arquitetura, não encargos operacionais permanentes, e a fricção que impõem recai desproporcionalmente sobre o pequeno número de classes de ação para as quais foram construídas, e não sobre o trabalho de rotina da frota. Esse é o mais próximo que este catálogo chega de uma recomendação geral: a cobertura structural é, pela evidência de seus próprios campos de custo e fricção, mais barata de construir do que comumente se presume, e a camada instructional que já existe a custo quase zero não é motivo para adiá-la, porque o custo quase zero é exatamente o motivo pelo qual a camada instructional não reduziu a lacuna de cobertura que a Seção 3 documenta.

8. Os dez condition sets experimentais

Um condition set é um subconjunto nomeado e explícito deste catálogo, usado para manter a variável guardrail constante em um nível de maturidade definido, de modo que seu efeito sobre autonomia e supervisão possa ser medido, e não argumentado. Dez são definidos, `CS-01` a `CS-10`, de uma linha de base nua com zero guardrails ativos até uma configuração totalmente endurecida contendo todo controle catalogado, mais uma ablação deliberada. A pertença é explícita por conjunto, e não inferida a partir de population ou camada, o que significa que a presença de um controle em um condition set é, ela própria, uma decisão registrada, rastreável em `condition-sets.json`, e não um artefato de como este catálogo por acaso está organizado.

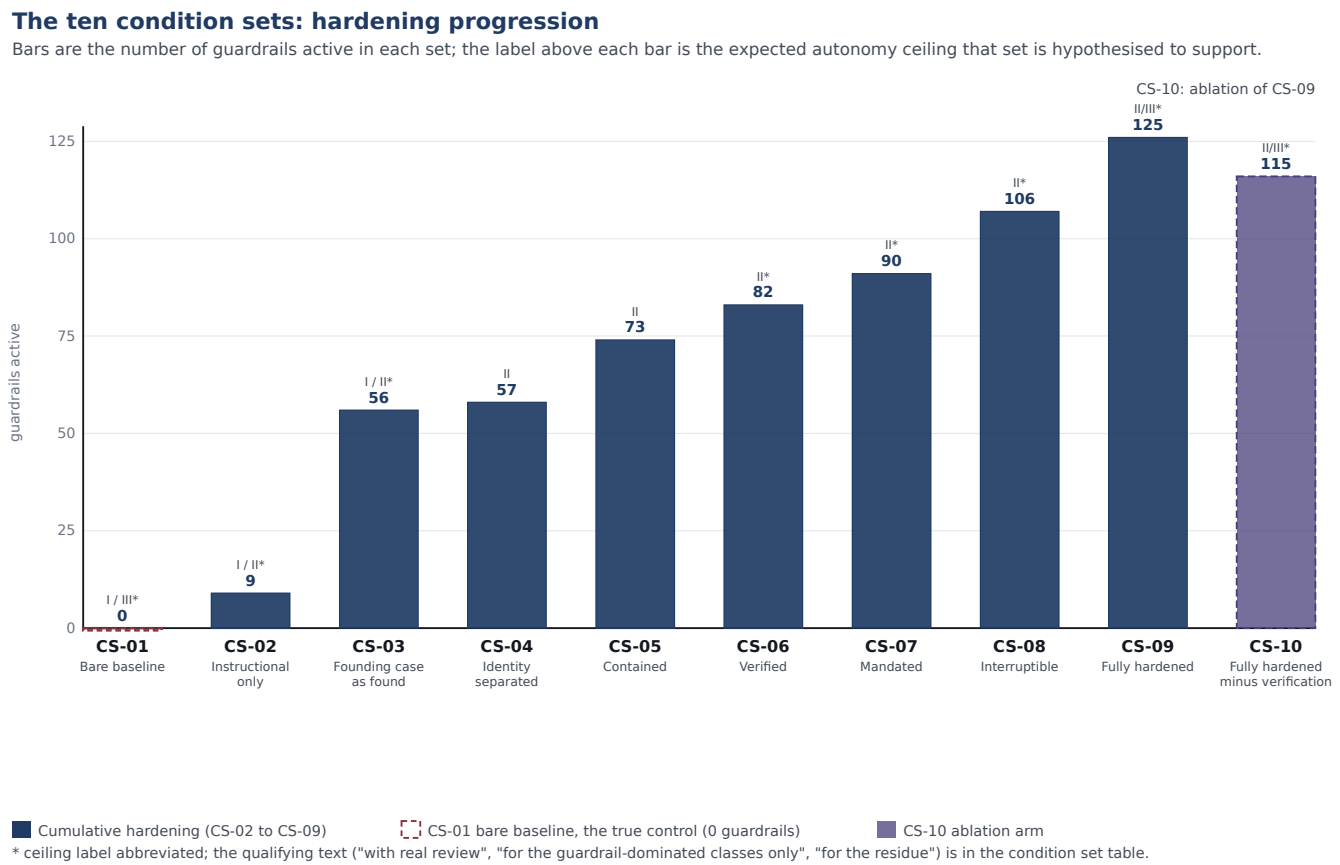

Figura D3-2. Os dez condition sets ordenados pela contagem de guardrails, de zero em CS-01 a 125 em CS-09. O CS-10 é desenhado como uma ablação do CS-09, e não como um passo adicional na progressão, porque ele remove, em vez de acrescentar, em relação ao conjunto imediatamente anterior.

Tabela D3-3. Os dez condition sets experimentais, da linha de base nua até a totalmente endurecida.

ID	NAME	PURPOSE	LAYERS COVERED	GUARD RAILS	EXPECTED CEILING	COST
CS-01	Bare baseline	The control condition. Establishes what the workload does when nothing bounds it, so that every other condition set can be read as a difference from it.		0	I in practice, III in principle for every class	none
CS-02	Instructional only	Isolates the contribution of instruction-based governance, which is the position the doctorate's central claim is directed against.	L12	9	I in practice, II claimed	very low
CS-03	Founding case as found	Anchors the experimental design to a real environment, so that the isolated environment can be shown to be a fair abstraction of one rather than a convenient fiction.	L12	56	I in practice, II claimed	low
CS-04	Identity separated	Tests whether identity separation alone, with no other new control, moves the boundary metrics. Identity is the layer on which every other layer depends.	L1, L12	57	II	moderate
CS-05	Contained	Tests the containment hypothesis in isolation: whether bounding what an error can reach is sufficient without bounding what the agent may claim.	L1, L2, L3, L12	73	II	moderate
CS-06	Verified	Tests the verification hypothesis, which is the founding case's own central finding turned into an intervention.	L1, L2, L3, L8, L12	82	II with real review	moderate
CS-07	Mandated	Tests whether the taxonomy made executable changes behaviour, as opposed to describing it after the fact.	L1, L2, L3, L7, L8, L12	90	II, with Level I defensible for the guardrail-dominated classes	high
CS-08	Interruptible	Tests the supervision viability conditions directly: whether the effective posture can be brought to match the nominal posture.	L1, L2, L3, L7, L8, L9, L10, L11, L12	106	II, with Level I defensible for the guardrail-dominated classes	high
CS-09	Fully hardened	Establishes the upper bound of the catalogue. What remains at Level II or III under this set is the irreducible supervision residue by construction.	L1, L2, L3, L4, L5, L6, L7, L8, L9, L10, L11, L12	125	II or III for the residue; I for everything else	very high
CS-10	Fully hardened minus verification	An ablation. Isolates the contribution of the verification layer by removing it from an otherwise maximal configuration, which is a stronger test than adding it to a minimal one.	L1, L2, L3, L4, L5, L6, L7, L9, L10, L11, L12	115	II or III for the residue; I for everything else	very high

Os nove conjuntos de `CS-01` a `CS-09` são ordenados pela contagem cumulativa de guardrails e por dependência, não por uma numeração arbitrária. `CS-01` **linha de base nua é o verdadeiro controle experimental**, zero guardrails ativos, explicitamente sinalizado `is_control` no registro, e seu teto esperado é declarado honestamente como dividido: Nível I na prática, porque nada impede um agente sem supervisão de simplesmente agir, contra Nível III em princípio para toda classe de ação, porque nada conquistou, para nenhuma delas, um requisito menor tampouco. `CS-02` **apenas instructional** acrescenta as nove entradas presentes da L12 e nada mais, o que faz dele o conjunto que isola exatamente o que a governança em prosa, isoladamente, tem a hipótese de alcançar; seu teto esperado, Nível I na prática e Nível II apenas reivindicado, é o condition set que este catálogo espera que pareça quase indistinguível de `CS-01` em qualquer métrica que meça a restrição real, e não a política declarada. `CS-03` **caso fundador como encontrado** não é uma configuração hipotética; é a

própria population presente-mais-parcial deste catálogo, cinquenta e seis guardrails, e sua reivindicação de teto é idêntica, em redação, à de `CS-02`, pelo mesmo motivo que a Seção 4 do D2 dá longamente: os controles conditional adicionais presentes no caso fundador mudam o que é verificado, não o que é alcançável de modo structural.

`CS-04` identidade separada é o primeiro conjunto a acrescentar cobertura structural genuinamente nova, em vez de mais verificação conditional, e seu teto esperado se move claramente para o Nível II, consistente com a hipótese da Seção 4 de que a L1 é necessária, mas não, por si só, suficiente para o Nível I. **`CS-05` contido** e **`CS-06` verificado** acrescentam, cada um, cerca de quinze a vinte e cinco guardrails adicionais, e espera-se que cada um sustente o Nível II, e não cruze para o Nível I, sendo o `CS-06` especificamente nomeado como alcançando revisão real, e não nominal, o que faz dele o condition set que este catálogo espera que mais mova a `MG-022` taxa de divergência de alegações, já que é o conjunto que instala as entradas de verificação independente da L8. **`CS-07` mandatado** e **`CS-08` interrompível** acrescentam, respectivamente, a maquinaria de política da L7 e a maquinaria de contenção da L9, e ambos carregam o teto qualificado de Nível II em geral, com o Nível I se tornando defensável especificamente para as classes de ação dominadas por guardrail que a Seção 2 do D2 identifica, o que é o teste mais claro, em toda a progressão de condition sets, da distinção entre `floor` e `ceiling` que aquele documento desenvolve: espera-se, como hipótese, que esses dois conjuntos movam o termo do teto sem mover o termo do piso, de forma alguma.

`CS-09` totalmente endurecido contém 125 das 134 entradas catalogadas. As nove não incluídas são oito das nove entradas classificadas `none` da Seção 1.2, todas exceto `GR-L10-07` visão de terminal ao vivo, que é passiva e nada enfraquece, além da única entrada instrucional `GR-L07-09` brief em prosa, que o conjunto endurecido substitui pela camada de mandato verificável por máquina. `GR-L01-01` conta de sistema operacional compartilhada está entre as oito excluídas e aparece apenas no `CS-03`, porque a partir do `CS-04` a camada de identidade fornece identidade por agente (`GR-L01-02`) em seu lugar. Seu teto esperado, Nível II ou III para o resíduo e Nível I para tudo o mais, é o condition set contra o qual o registro de resíduo do D2 é definido: uma classe de ação está no resíduo irredutível se, e somente se, permanecer acima do Nível I sob o `CS-09`.

`CS-10` totalmente endurecido menos verificação é a única ablação deliberada do conjunto, 115 guardrails, construída removendo do `CS-09` as entradas de verificação independente da L8 e nada mais. Ela existe para responder a uma pergunta específica, e não para ocupar mais um passo na progressão de endurecimento: quanto do efeito do `CS-09` sobre a divergência de alegações e sobre o resíduo é de fato atribuível à verificação independente, em contraste com a maquinaria de contenção, identidade e mandato que toda outra camada contribui. O D2 nomeia isso diretamente ao identificar `SD-05` não verificabilidade como o determinante mais responsivo a investimento, e a camada de verificação como o bloco isolado mais consequente do catálogo; `CS-10` é a ablação construída para testar se essa afirmação é correta, e não apenas plausível, e o fato de seu teto esperado ser declarado de forma idêntica ao de `CS-09`, Nível II ou III para o resíduo, Nível I para tudo o mais, é, por si só, uma previsão testável: se os dois conjuntos divergirem substancialmente na `MG-022` em uma execução real, a verificação está fazendo mais trabalho do que o teto de autonomia isoladamente sugeriria, porque teto e divergência são medidas diferentes do mesmo conjunto de guardrails.

> **CONDITION SETS SÃO UMA VARIÁVEL, NÃO UM VEREDITO**
>
> Nenhum condition set nesta tabela foi exercido em uma execução ao vivo contra o caso fundador ou contra o ambiente experimental doutoral. Todo teto esperado na tabela acima é uma hipótese, enunciada como tal no próprio campo `hypothesis` do registro, e o propósito de nomear dez conjuntos explicitamente, em vez de descrever a maturidade dos guardrails como uma única escala contínua, é permitir que uma execução posterior relate qual hipótese específica se sustentou e qual não, entrada por entrada, em vez de relatar uma impressão de se o endurecimento em geral pareceu ajudar.