

Lié Localement. Gouverné Globalement.

Gestion du Risque Systémique pour les Systèmes Probabilistes Distribués

Le Problème Distinctif

La question que nous adressons ne concerne pas l'autonomie de l'agent per se. Elle concerne la propagation d'assertions incertaines à travers une autorité décisionnelle déléguée vers des effets du monde réel qui interagissent.

C'est important parce qu'un système peut échouer même quand aucun composant ne viole son contrat local. Un agent SafetyTriage reste dans ses seuils de confiance. MedicalReview reste dans ses limites d'escalade. RegulatoryReporting reste dans son autorité de dépôt. Pourtant le système produit un résultat qu'aucun d'eux n'avait prévu : escalades en cascade qui submergent la capacité, ou erreurs corrélées qui s'amplifient dans la flotte, ou boucles de rétroaction qui déstabilisent l'ensemble.

Le problème n'est pas des agents défectueux. Le problème est que les systèmes probabilistes distribués exhibent une émergence. Les décisions locales interagissent. Les assertions se composent. Les effets s'accumulent. Le système devient une entité d'ordre supérieur avec des propriétés non réductibles aux propriétés des composants.

Ceci est la thèse : les systèmes agentiques exigent une gouvernance systémique non parce que les agents sont peu fiables, mais parce que l'émergence est réelle. Le confinement local est nécessaire mais insuffisant.

Plus précisément : la gouvernance systémique est un problème de contrôle en boucle fermée sous observabilité partielle. Le plan de contrôle ne peut pas observer parfaitement le système qu'il régule. Il fonctionne à partir d'observations bruitées, construit des estimations d'état, et prend des décisions basées sur des estimations qu'il sait incomplètes. La rétroaction détermine si le système se stabilise ou diverge.

Les Catégories de Risque

Pour comprendre pourquoi la sécurité locale échoue systématiquement, nous devons partitionner le risque attentivement.

Le risque local est la ligne de base : un agent fait une inférence probabiliste et n'est correct que avec probabilité p . Les hallucinations de modèle, les biais d'entraînement, les décalages de distribution, la contamination de données : ce sont des propriétés locales.

Les mécanismes de propagation convertissent le risque local en risque systémique par des canaux distincts.

Le premier est l'amplification : les assertions portent l'incertitude épistémique, mais les agents aval les traitent comme des faits. Un SafetyModel génère « risque élevé » avec confiance 0,65, marqué INFÉRÉ. Le Safety Committee voit ceci et escalade avec confiance 0,78. Par RegulatoryReporting, c'est traité comme fait OBSERVÉ avec confiance 0,85. Le statut épistémique original a été blanchi. La fausse certitude pilote le système.

Le deuxième est la corrélation : si tous les agents dépendent du même modèle, tous les agents amplifient la même erreur simultanément. Ce n'est pas l'indépendance. C'est la défaillance synchronisée dans une flotte supposément diverse.

L'interaction est un troisième phénomène systémique, distinct de la propagation. La décision d'un agent change le contexte pour le suivant, non par amplification d'erreur, mais par changement de l'environnement. L'Agent A modifie un score de priorité. L'Agent B observe le nouveau score et réalloue les ressources. L'Agent C voit la capacité disponible et escalade une décision qu'il aurait précédemment retenue. Aucun n'excède individuellement l'autorité. Ensemble ils forment une boucle de rétroaction que personne n'a causée. C'est le système qui l'a fait.

L'interaction diffère de l'amplification (pas de fausse confiance) et de la corrélation (pas de modèle partagé). C'est l'émergence pure : l'optimisation locale produit la dysfonction globale par rétroaction couplée.

Inversement, les mécanismes de résilience systémique peuvent réduire ou contenir le risque : Diversification (les agents dépendent de modèles différents), Redondance (les fonctions critiques ont des secours), Exigences de quorum (les escalades exigent l'accord), Dégradation gracieuse (le système réduit la fonctionnalité plutôt que de s'effondrer), Mécanismes de récupération (arrêt actif et réinitialisation d'état).

Les facteurs structurels déterminent si la propagation et l'interaction deviennent catastrophiques. Concentration de dépendance : combien de domaines de défaillance se cachent derrière la diversité apparente ? Saturation de capacité : le système peut-il absorber les effets cumulatifs ? Incertitude d'estimation d'état : le plan de gouvernance sous-estime-t-il ou surestime-t-il l'état ?

Le temps est la dimension dans laquelle tout évolue. Cent escalades par mois différent de cent par seconde. Le temps de récupération du système compte. Le système est-il entré dans un régime plus difficile à inverser ?

Formule de risque : $R_{\text{système}} = F(R_{\text{local}}, \text{Amplification}, \text{Corrélation}, \text{Interaction}, \text{Mécanismes_Résilience}, \text{Facteurs_Structurels}, \text{Dynamiques_Temporelles})$

F est une fonction de dépendance conceptuelle, non additive. La relation dépend de l'architecture.

Assertion, Décision, Effet

L'Article 1 a établi une chaîne : la Capacité s'écoule vers l'Exécution, produisant des Assertions, pilotant les Décisions, générant les Effets. Cet article gouverne cette chaîne à chaque étage.

La gouvernance d'assertion prévient l'effondrement de type épistémique et préserve la provenance. Une assertion porte deux propriétés : le contenu (ce que nous avons appris) et le statut épistémique (comment nous le savons). Une inférence d'un modèle a le statut « (Modèle, Inféré, T_valide) ». Une observation d'un humain a le statut « (Humain, Observé, T_valide) ». Ceux-ci sont orthogonaux au contenu.

Le problème est que les systèmes aval jettent le statut épistémique. Un SafetyModel produit « Risque=ÉLEVÉ, confiance=0,65, statut_épistémique=INFÉRÉ, valide_jusqu'à=T+10min ». Au moment où cela atteint la Décision, seul « Risque=ÉLEVÉ » subsiste. Le système croit que c'est une observation.

Pire : les assertions sont souvent transformées, non simplement relayées. Une inférence SafetyModel produit une recommandation. Un Safety Committee fusionne les recommandations multiples et produit une escalade. L'escalade n'est pas simplement l'inférence du modèle avec statut attaché. C'est une synthèse. Quel est le statut épistémique de la synthèse ?

La gouvernance d'assertion applique la préservation de type aux limites de service et trace la provenance épistémique. Une règle énonce : « (Modèle, Inféré) ne peut pas devenir (Humain, Observé) sans validation explicite. » Quand la synthèse se produit, la provenance est enregistrée comme une chaîne : quels modèles ont contribué ? Quels humains ont examiné ? Avec quelle confiance ? Le système maintient un graphe acyclique orienté d'origine pour chaque assertion.

Ce n'est pas une perte de données. C'est l'intégrité des données au niveau sémantique.

La gouvernance de décision applique les limites d'autorité et la réponse proportionnée sous incertitude. La même assertion sous une autorité différente produit un risque différent. Une recommandation avec autorité « proposer seulement » porte un risque bas ; l'humain peut refuser. La même recommandation avec autorité « exécuter » porte un risque élevé.

La gouvernance de décision applique l'intervention proportionnée quand l'incertitude est élevée. Quand un signal apparaît (le taux d'escalade diffère entre démographies), la réponse dépend de : (solidité de la preuve, préjudice potentiel si faux, coût de l'intervention, réversibilité, informatif).

Exemple : les taux d'escalade diffèrent 2x entre populations. Le préjudice potentiel (iniquité du patient) est élevé. Le coût de réduire l'autorité d'escalade est bas. La réduction est facilement réversible si les modèles changent.

Par conséquent : la réduction d'autonomie conservatrice est justifiée même sans preuve causale. C'est le test d'hypothèse, pas de la recklessness.

La gouvernance d'effet trace les conséquences à deux niveaux. La réversibilité locale demande : si cette décision est fausse, peut-elle être défaite ? L'irréversibilité globale demande : si N telles décisions vont toutes mal, le système peut-il se récupérer ? Cent escalades SafetyCase, chacune localement réversible, peuvent collectivement submerger la file d'attente d'examen. Le système entre un état où aucune quantité de dé-escalade ne le récupère parce que les nouveaux cas arrivent plus vite que la capacité d'examen.

C'est l'irréversibilité au niveau du système. La gouvernance locale la manque.

Gouvernance de l'État du Système

Le quatrième objet de gouvernance est l'état du système. Ce n'est pas un « plan de gouvernance ». C'est la condition inférée du système, déterminée par les effets cumulatifs et agrégée à partir des observations.

L'État du Système est fondamentalement différent de l'Assertion, la Décision, et l'Effet. Les trois derniers sont des événements observables. L'État du Système est inféré à partir d'événements. Il porte sa propre incertitude épistémique.

Le plan de gouvernance n'observe pas directement l'état. Il observe les décisions et les agrège : « Au cours de la dernière heure, 47 escalades se sont produites, 23 modifications ont été exécutées, 12 inversions ont été demandées ». À partir de ces observations, il construit une estimation utilisant un modèle d'espace d'état :

État_estimé_t = g(Observations_t, État_estimé_t-1, Méthode_Agrégation)

Le plan de gouvernance agit sur État_estimé, non sur État_vrai.

Ceci introduit une exigence fondamentale : l'observabilité. L'état vrai du système peut-il être reconstruit à partir des sorties observables ? Sinon, le plan de gouvernance est aveugle. L'identifiabilité d'état est une précondition pour un contrôle efficace.

Formellement : le système est observabilité-état si pour deux états vrais distincts S1 et S2, il existe une séquence d'observations qui les distingue. C'est le fondement mathématique de pourquoi l'instrumentation compte. Sans instrumentation suffisante, la gouvernance devine littéralement.

La rétroaction naît naturellement de cette structure. Si la gouvernance réduit l'autonomie parce que la capacité estimée est basse, la capacité réelle s'améliore. Mais l'estimation peut prendre du retard. La gouvernance conservatrice peut créer une contrainte inutile. La rétroaction peut causer du dépassement.

Trois Enveloppes Distinctes

Les cadres actuels fusionnent trois concepts différents sous « autonomie ». Nous devons les séparer.

L'Enveloppe d'Autonomie décrit ce qu'un agent est autorisé à faire. C'est un espace d'action contraint par la politique. L'agent peut lire n'importe quelles données, recommander l'escalade jusqu'à 10 par heure, ne peut pas exécuter les modifications sans approbation humaine. L'enveloppe est multidimensionnelle : type d'action, périmètre, limites de ressources, règles de contexte.

L'Enveloppe de Capacité décrit la capacité du système à absorber et traiter les effets sans entrer dans des états non récupérables. Ce n'est pas un simple nombre. C'est une propriété dynamique :

$$\text{Capacité_disponible}(t) = \text{Taux_Service}(t) - \text{Charge}(t) - \text{Backlog}(t)$$

où :

- Taux_Service(t) : à quelle vitesse le système traite les escalades, modifications, examens
- Charge(t) : taux d'arrivée des nouvelles demandes
- Backlog(t) : demandes non traitées des périodes antérieures

La capacité est épuisée quand Charge > Taux_Service et Backlog croît. La récupération se produit quand Taux_Service > Charge assez longtemps pour vider le Backlog. Le temps de récupération dépend de la profondeur du backlog.

Le Budget de Risque est l'allocation spécifique au domaine de la perte tolérable. Combien d'escalades par jour ? Combien d'exposition financière ? Combien d'incidents réglementaires par trimestre ?

Les budgets de risque sont par domaine parce que les effets sont incommensurables. Vous ne pouvez pas ajouter escalades cliniques et expositions financières en un seul score numérique. Mais vous pouvez les suivre en parallèle, avec des seuils séparés.

La relation :

Capacité_disponible(t) contraint Budget_Risque_disponible(t)

Budget_Risque_disponible(t) contraint Enveloppe_Autonomie(t)

L'Action est autorisée si :

**** Politique_Autorise(action) ET****

**** Coût_Risque(action) <= Budget_Risque_restant[domaine] ET****

**** Taux_Service_disponible >= Taux_Service_requis(action)****

Trois enveloppes, trois mécanismes distincts, trois canaux de rétroaction séparés.

Estimation d'État Sous Observabilité Partielle

L'insight récuratif : le plan de gouvernance produit des assertions sur l'état du système. Ces assertions portent l'incertitude épistémique. Le plan de contrôle est sujet aux mêmes limites épistémiques qu'il impose aux agents.

Quand le plan de gouvernance estime « Système à 87% d'utilisation de capacité », c'est une assertion. Elle porte la confiance (0,73) et la fenêtre de validité (5 minutes). Mais le plan de gouvernance la traite comme un fait quand il prend des décisions. C'est exactement le problème que nous avons diagnostiqué au niveau de l'agent.

Le plan de gouvernance doit être transparent sur son propre statut épistémique. Les estimations d'état portent des limites explicites. Les décisions de contrôle sont prises sachant l'incertitude.

Quand l'incertitude est élevée, quelle réponse de gouvernance est appropriée ? La réponse dépend du contexte : si l'action est irréversible ou coûteuse, le conservatisme est justifié. Si l'action est réversible et informative (réduit l'incertitude), l'exploration peut être justifiée malgré l'incertitude élevée.

La formulation correcte : la gouvernance adapte le conservatisme basé sur (incertitude, réversibilité, informatif, coût).

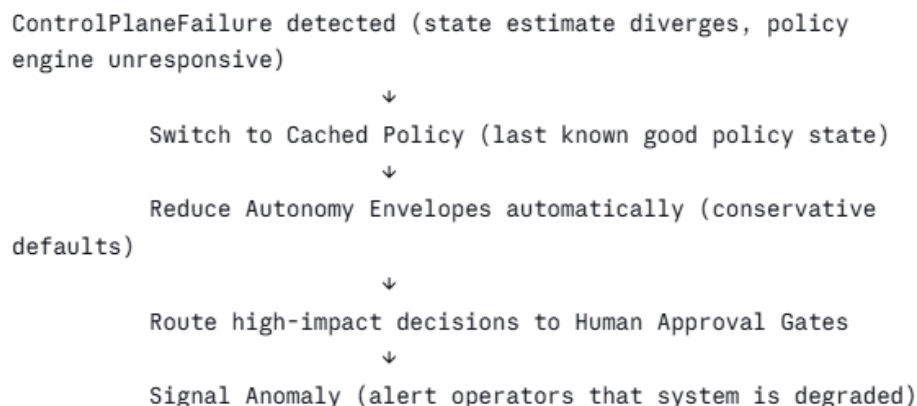
La gouvernance structurellement adverse au risque est sous-optimale. Elle maximise la contrainte aux dépens de la connaissance.

Défaillance du Plan de Contrôle : Le Mode Dégradé

Voici la lacune architecturale critique : que se passe-t-il si le plan de contrôle lui-même échoue ?

Si le moteur de politique s'écroule, si l'estimateur d'état diverge, si l'autorité de décision est compromise, le système ne doit pas s'effondrer dans la paralysie ou l'action incontrôlée.

Le système exige un mode dégradé :



En mode dégradé :

- Les agents opèrent sous une politique gelée, pas une politique vivante. La politique depuis le dernier état d'estimation réussi est utilisée. Pas de mise à jour de politique jusqu'à la récupération du plan de contrôle.
- Les enveloppes d'autonomie se contractent. Où la politique autorisait 50 escalades par heure, le mode dégradé en autorise 10. Où la politique autorisait la modification de n'importe quel enregistrement, le mode dégradé autorise seulement l'accès en lecture.
- Les actions à haut impact (escalades affectant la sécurité du patient, modifications affectant les données cliniques, dépôts réglementaires) acheminent vers l'approbation humaine. Le système est temporairement médiatisé par les humains.

Ce n'est pas une défaillance. C'est une dégradation gracieuse. Le système perd l'optimisation mais conserve la sécurité.

Le plan de contrôle lui-même doit être conçu comme un système tolérant les pannes avec des minuteries de surveillance, des mécanismes de secours, et une détection explicite de défaillance. Le mode dégradé n'est pas un espoir. C'est une architecture. Sans lui, le plan de contrôle est une défaillance de mode commun pour tout le système.

C'est le dernier morceau architectural qui ferme la conception. La gouvernance n'est pas externe au système. C'est une infrastructure critique qui peut échouer. Cet échec doit être anticipé et atténué par conception.

Agents Liés. Gouvernance Exposée.

Cette phrase résume le changement de paradigme. Nous lions l'autorité locale et redirigeons la gouvernance vers l'exposition au niveau du système.

La gestion du risque financier fournit des modèles de contrôle réutilisables pour ce problème. Les limites de position contraignent les traders individuels. Les limites de portefeuille contraignent la flotte. Les limites de concentration détectent la monoculture. Le test de stress évalue la stabilité du système. Les disjoncteurs arrêtent la propagation. Les protocoles de récupération permettent l'arrêt gracieux.

Les systèmes agentiques peuvent réutiliser ces modèles :

- Validation pré-décision (vérification de statut épistémique, vérification d'autorité)
- Surveillance post-décision (la décision était-elle optimale ? A-t-elle consommé les budgets comme prévu ?)
- Analyse de scénario (si le modèle dérive, le système est-il stable ?)
- Interrupteurs d'arrêt (arrêter l'escalade si l'état se dégrade)
- Modes de récupération (arrêt gracieux, passage à l'opération dégradée)

Le changement conceptuel : de « cet agent est-il sûr ? » à « le système absorbe-t-il le risque au taux que nous pouvons tolérer ? »

Le Paradigme : Recalibration Continue

La gouvernance n'est pas un niveau de maturité à atteindre. C'est un point opérationnel à maintenir. La gouvernance optimale existe, mais elle n'est ni unique ni stable.

$$G_t = \operatorname{argmin} [\operatorname{ExpectedLoss}_t(G) + \operatorname{GovernanceCost}_t(G)]$$

C'est un modèle conceptuel, non directement calculable. Mais le principe tient : la gouvernance opère à un optimum qui se déplace quand les conditions changent. Les cadres qui résistent à la recalibration deviennent des obstacles.

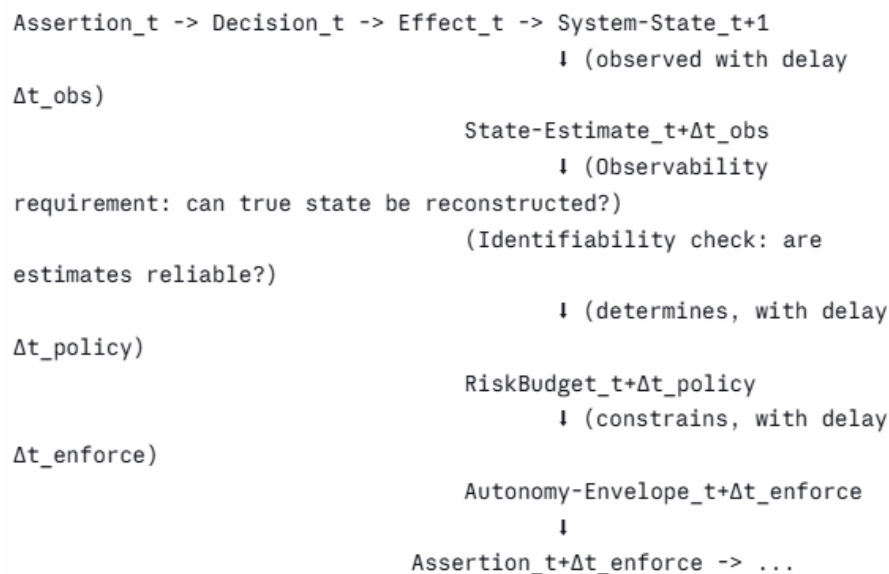
La recalibration continue exige :

- État observable (instrumentation adéquate)
- Flexibilité de politique (capacité à ajuster les contraintes rapidement)
- Boucles de rétroaction (les conséquences retournent à la gouvernance)
- Seuils adaptatifs (plutôt que des règles fixes)

C'est la réponse proportionnée aux conditions changeantes, pas du hasard.

La Cascade avec Délais et Identifiabilité

Le système opère comme une boucle de rétroaction avec délais de propagation.



Les délais importent. Le délai d'observation signifie que l'estimation prend du retard sur la réalité. Le délai de politique signifie que les contraintes sont basées sur des informations obsolètes. Le délai d'application signifie que de nouvelles assertions sont faites avant que les anciennes contraintes prennent effet.

L'instabilité peut naître non pas d'une mauvaise politique mais de délais dépassant le temps d'intégration du système. La théorie du contrôle enseigne : les délais dans les boucles de rétroaction négative peuvent causer l'instabilité et l'oscillation.

Plus fondamentalement : si le système n'est pas observable-état, la boucle de rétroaction est bâtie sur du contrôle aveugle. La gouvernance ne peut pas reconstruire ce que le système fait réellement. La boucle devient réactive plutôt que prédictive.

L'identifiabilité d'état est le fondement. L'observabilité est l'exigence d'ingénierie. Les deux doivent être conçues avant que la gouvernance puisse être efficace.

Résilience Systémique et Pathologie

L'émergence est neutre. Elle peut être protectrice ou pathologique.

L'émergence protectrice : la redondance prévient la cascade, les exigences de quorum préviennent le biais de modèle unique, la dégradation gracieuse préserve la fonction centrale. Ce sont des mécanismes de résilience systémique.

L'émergence pathologique : l'amplification en cascade, les défaillances synchronisées, les boucles de rétroaction qui déstabilisent, la surgouvernance qui paralyse.

La tâche de gouvernance est d'améliorer l'émergence protectrice et de supprimer l'émergence pathologique. Vous ne pouvez pas prévenir l'émergence. Vous pouvez seulement la façonner.

La gouvernance optimale équilibre : contrainte suffisante pour prévenir l'émergence pathologique, liberté suffisante pour permettre l'émergence protectrice.

Fermeture de la Trilogie

La trilogie se ferme maintenant sur elle-même.

L'Article 1 demandait : quelle est la primitive ? Réponse : Capacité. L'agent est un choix d'implémentation.

L'Article 2 demandait : comment savons-nous ? Réponse : Les assertions portent statut épistémique, orthogonal au contenu. L'autorité est indépendante de l'épistémologie.

L'Article 3 demande : comment contrôlons-nous ? Réponse : Le plan de contrôle lui-même est un sous-système sujet aux mêmes limites épistémiques qu'il impose aux agents. Il doit être conçu comme une infrastructure tolérant les pannes.

Aucun gouverneur n'a un accès privilégié à la vérité de base. La gouvernance elle-même opère par observation, inférence, et intervention, tous intégrés dans le système gouverné.

L'architecture qui émerge :

Autorité locale liée. Les enveloppes d'autonomie sont contraintes par la politique et la capacité.

Provenance épistémique préservée. Les assertions portent des chaînes d'origine, pas seulement du contenu.

Exposition du système gouvernée. Les budgets de risque contraignent les effets cumulatifs, pas seulement les actions individuelles.

Boucle de rétroaction contrôlée. L'estimation d'état, la décision de politique, et l'application retournent tous vers le système qu'ils régulent.

Et la précondition : le système doit être observabilité-état et le plan de contrôle doit être tolérant aux pannes avec des modes dégradés explicites.

C'est plus profond que « mieux gérer les agents ». C'est l'architecture du contrôle pour les systèmes probabilistes distribués sous observabilité partielle.

Domaine de Validité et Périmètre

Ce cadre s'applique avec la plus haute valeur aux systèmes où :

- Le risque est multidimensionnel (conséquences cliniques, financières, réglementaires)
- Les modes de défaillance sont émergents (pas des erreurs au niveau des composants)
- La récupération est dépendante du temps (le système ne peut pas se réinitialiser instantanément)
- Les boucles de rétroaction sont significatives (les changements d'état affectent les décisions de gouvernance)
- L'observabilité partielle est inhérente (l'état du système ne peut pas être parfaitement connu)

Il s'applique avec une valeur plus basse aux systèmes où :

- Le risque est unidimensionnel et aisément quantifiable
- Les modes de défaillance sont principalement au niveau des composants
- La récupération est instantanée (opération sans état)
- Les boucles de rétroaction sont négligeables

- L'observabilité complète est réalisable

Le cadre ne prétend pas à l'universalité. Il est optimisé pour les domaines régulés où l'émergence compte et où l'auditabilité est requise : santé, finance, déploiement d'IA régulé.

Exemple : Le Terrain PREDICARE

PREDICARE (médecine prédictive territoriale, France 2030) illustre ces principes. Les agents incluent SafetyTriage (lit les signaux de sécurité, recommande l'escalade), MedicalReview (conduit par humain, autorité d'escalader), RegulatoryReporting (autorité de dépôt).

Les budgets de risque sont définis :

- Escalades cliniques : 100 par jour avant déclenchement de l'examen manuel
- Modifications des données patient : 1000 par jour avant verrouillage du système
- Dépôts réglementaires : 1 par trimestre avant escalade

Les enveloppes d'autonomie sont dynamiques, liées à la capacité :

- Quand la capacité d'escalade est en-dessous de 20% disponible, réduire l'autonomie d'escalade
- Quand le débit de modification excède 80% de la capacité, bloquer les modifications non urgentes

Le système exhibe une rétroaction protectrice. Serrer les contraintes quand la capacité est épuisée, relâcher quand la capacité se récupère.

C'est illustratif de l'architecture, non une prétention sur la performance mesurée. PREDICARE reste en déploiement. La validation quantitative exige la méthodologie complète et les données.

La défaillance du plan de contrôle est atténuée : si le moteur de politique en temps réel échoue, le système revient à une politique gelée et réduit l'autonomie aux défauts conservateurs. Les portes d'approbation humaine deviennent actives pour les décisions à haut impact.

Ce Que Cela Permet

Le système opère maintenant sous les principes qui sont :

- **Distribué** : Aucun observateur central n'a l'état complet. Le contrôle émerge de la rétroaction locale.
- **Fondé en principes** : Les contraintes dérivent des budgets de risque et des limites spécifiques au domaine.
- **Adaptatif** : Les points opérationnels se recalibrent quand les conditions changent.

- **Observable** : Les décisions sont justifiées par les estimations d'état et la comptabilité du risque.
- **Tolérant aux pannes** : La défaillance du plan de contrôle déclenche une dégradation gracieuse, non un effondrement.

Conclusion : Le Tournant Théorique du Contrôle

La trilogie se ferme avec une reconnaissance qui n'est devenue explicite qu'à cette itération finale.

La gouvernance systémique est un problème de contrôle en boucle fermée sous observabilité partielle. Le plan de contrôle n'est pas séparé du système contrôlé. C'est un sous-système réflexif imbriqué en lui, sujet aux mêmes limites épistémiques qu'il impose.

Le plan de gouvernance ne peut pas sortir du système pour le gouverner. Il ne peut que réguler les boucles qui le constituent. Et le plan de gouvernance lui-même peut échouer. Cet échec doit être anticipé par conception.

La contribution centrale : les propriétés des composants et les propriétés du système sont distinctes.

Appliqué :

- La capacité locale et l'architecture systémique ne sont pas la même chose
- Le statut épistémique de l'assertion et la valeur de vérité de l'assertion sont orthogonaux
- La réversibilité locale et la réversibilité systémique ne sont pas équivalentes

L'architecture qui adresse ceci : Autorité locale liée. Provenance épistémique préservée. Exposition du système gouvernée. Boucle de rétroaction contrôlée. Plan de contrôle tolérant aux pannes.

Bâtir en conséquence.