La gouvernance agentique ne viendra pas des modèles

Action-space, autonomie, réversibilité et régimes de décision : pour une architecture de gouvernance exogène des systèmes d'IA agentique

Jérôme Vetillard / Twingital Institute® Avril 2026

Résumé

Les systèmes d'IA agentique atteignent en 2026 un seuil de déploiement industriel qui rend urgente la question de leur gouvernance. Cet article soutient que la gouvernabilité d'un système agentique ne peut pas être attendue comme une propriété émergente du seul progrès des modèles. Elle doit être produite par l'architecture du système qui les met en situation d'agir. L'article propose une grille de risque structurée autour de trois axes :

- L'étendue de l'action-space,
- Le degré d'autonomie,
- La réversibilité des actions,

complétée par deux modulateurs, la criticité du domaine et l'asymétrie d'erreur. Cette grille fonde une taxonomie de trois régimes de décision :

- assistance,
- recommandation structurée,
- exécution bornée,

qui ne forment pas une hiérarchie de maturité mais des modes architecturaux distincts. L'article examine ensuite le pattern des contrats de décision, dérivé de la logique du design-by-contract, comme mécanisme opérationnel de contrôle d'admission. Les cadres institutionnels émergents de Singapour, des États-Unis, de l'Union européenne et du Royaume-Uni sont analysés. Six limites et un programme de recherche en quatre questions sont explicitement formulés.

1. Le problème : des agents plus capables, pas gouvernables par construction

Le problème central de l'IA agentique n'est pas que les agents deviennent plus capables. C'est qu'ils deviennent actionnables avant d'être gouvernables.

Les systèmes d'IA agentique atteignent en 2026 un seuil de visibilité industrielle qui impose un examen architectural spécifique. Les analystes de marché anticipent une diffusion rapide d'agents task-specific dans les applications d'entreprise au cours des prochaines années. Gartner projette ainsi que jusqu'à 40 % des applications d'entreprise intégreront des agents task-specific d'ici la fin de l'année 2026, contre moins de 5 % en 2025. Cette estimation relève d'une prévision de marché et non d'un constat empirique consolidé, mais elle signale une direction stratégique claire : l'agent n'est plus seulement une démonstration technique. Il devient un composant plausible, puis progressivement banal, des systèmes d'information contemporains.[1]

Cette montée en puissance ne résout pourtant rien de l'enjeu principal. Un système agentique n'est pas seulement un système qui produit du texte, du code ou une classification. C'est un système qui peut, dans certaines configurations, sélectionner des outils, mobiliser des données, déclencher des actions, enchaîner des opérations, réviser son plan et, parfois, agir sur des artefacts ou des environnements extérieurs au modèle lui-même. Le problème ne réside donc pas seulement dans la qualité de la sortie. Il réside dans le régime de décision introduit par l'architecture.

Le déplacement est décisif. Dans un workflow déterministe, le graphe de contrôle est défini à l'avance : un composant s'exécute à un endroit prévu, dans un ordre prévu, avec un périmètre d'effet relativement stable. Dans un système agentique au sens fort, le composant dispose d'une latitude partielle sur le choix des étapes, des outils ou des séquences d'action. ***Le système n'exécute plus seulement un chemin ; il explore un espace de décision***. Ce n'est pas une nuance terminologique. C'est un changement de régime. Et c'est précisément ce qui rend la gouvernance agentique distincte de la gouvernance logicielle classique.

Cette distinction est d'autant plus importante que les progrès récents de capacité ne se confondent pas avec des progrès équivalents de fiabilité. Les travaux de Rabanser, Kapoor, Kirgis, Liu, Utpala et Narayanan proposent un cadre d'évaluation structuré autour de quatre dimensions, la consistance, la robustesse, la prédictibilité et la sécurité, à partir de douze métriques concrètes. Leur intérêt n'est pas de démontrer une impossibilité théorique de rendre les agents fiables, mais de documenter un écart empirique : les gains de capacité observés ne se traduisent pas mécaniquement en gains substantiels et homogènes sur les dimensions de fiabilité les plus pertinentes en

contexte opérationnel. Autrement dit, le fait qu'un agent puisse davantage faire ne signifie pas encore qu'il fasse ce qu'il fait de manière stable, reconstructible et prédictible.[2]

Le présent article soutient, à partir de ce constat, une thèse simple. ***La gouvernabilité d'un système agentique ne peut pas être attendue comme une propriété émergente du seul progrès des modèles***. Elle doit être produite par l'architecture du système qui les met en situation d'agir. Cette gouvernance sera dite ici exogène non parce qu'elle ignorerait les propriétés internes du modèle, mais parce qu'elle ne dépend pas d'elles pour exister. Par gouvernance exogène, on désigne l'ensemble des mécanismes de contrôle, de validation, de traçabilité, de bornage et d'audit qui sont extérieurs au modèle lui-même : politiques déclaratives, contrôleurs d'admission, contrats de décision, checkpoints humains, séparation des droits d'action, allowlists, journaux d'évidence et mécanismes de révocation. La propriété de gouvernabilité n'est pas alors une qualité psychologique prêtée au modèle ; elle devient une propriété du système.

Il faut cependant éviter une dramatisation artificielle. ***Tout ce qui se présente aujourd'hui comme "agent" n'est pas un agent au sens fort***. Simon Willison décrit un paysage où les usages réellement utiles passent souvent par des patterns d'ingénierie cadrés et fortement instrumentés, notamment dans les agents de code, plutôt que par une autonomie ouverte et indifférenciée. Cette observation reste celle d'un praticien et non d'une enquête de marché. Mais elle a une portée analytique utile. Elle rappelle que le problème traité ici concerne moins l'ensemble des usages LLM en entreprise que la fraction de systèmes dans lesquels un espace décisionnel réel est accordé au composant. C'est précisément cette fraction, encore minoritaire en volume mais structurante en risque, qui concentre le plus fortement les enjeux de gouvernance.[3]

La question n'est donc pas de savoir si les modèles deviennent plus impressionnants. Elle est de savoir si un système qui leur délègue une part d'initiative peut rester opposable, explicable, révisable et responsable lorsque ses décisions ont des effets sur des tiers. C'est à ce niveau que la gouvernance cesse d'être un supplément de conformité pour devenir un problème d'architecture.

2. La contre-thèse : l'amélioration endogène des modèles rendra-t-elle les garde-fous architecturaux marginaux ?

La thèse d'une gouvernance exogène forte doit prendre au sérieux sa meilleure objection. Cette objection peut être formulée ainsi : les modèles progressent de génération en génération, les techniques d'alignement réduisent certaines formes d'écart comportemental, les sorties se structurent mieux, les outils de benchmarking se multiplient, et l'on peut envisager qu'à mesure que la fiabilité mesurée augmente, le

niveau d'autonomie concédé au système soit ajusté dynamiquement. Dans cette perspective, les dispositifs architecturaux lourds de gouvernance ne seraient qu'un état transitoire. Le système deviendrait progressivement gouvernable parce que son composant central deviendrait lui-même plus stable, mieux calibré, mieux aligné et mieux surveillé.

Cette position ne doit pas être caricaturée. Elle s'appuie sur une trajectoire empirique réelle. Les modèles récents sont, à bien des égards, plus utiles et plus structurés que leurs prédécesseurs : les system cards et évaluations publiées par plusieurs laboratoires documentent des gains mesurables en suivi d'instructions, en structuration des sorties et en gestion de formats contraints.[10] Il serait absurde de nier l'intérêt de ces progrès endogènes. La question n'est donc pas de les opposer dogmatiquement à toute gouvernance architecturale. La question est de savoir s'ils suffisent.

Trois raisons conduisent à répondre négativement.

La première tient à la distinction entre capacité et consistance. La capacité peut croître avec l'échelle, la qualité des données, la sophistication de l'entraînement ou des méthodes d'alignement. La consistance, au sens d'une stabilité de comportement sur des tâches similaires, dans des environnements proches, avec un régime d'erreur compréhensible et bornable, ne suit pas nécessairement la même dynamique. Les travaux de Rabanser et al. documentent précisément ce décalage. Ils n'établissent pas une impossibilité de convergence future, mais ils montrent qu'en l'état des systèmes observés, la progression de capacité ne permet pas encore de traiter la gouvernance comme un simple sous-produit du scaling.[2]

La deuxième raison est que tous les risques pertinents ne résident pas dans le modèle pris isolément. Le cadre AI TRiSM de Gartner vise la gouvernance, la fiabilité, la robustesse, la protection des données et la sécurité des déploiements IA. Lorsqu'il est prolongé dans la littérature consacrée aux systèmes multi-agents et agentiques, il éclaire des classes de risques systémiques qui ne se réduisent pas à la qualité intrinsèque d'un composant : cascades d'erreurs, interactions imprévues, contournement de mécanismes d'oversight, ou dégradation induite par des espaces de contexte et de mémoire partagés. Un composant localement satisfaisant peut donc produire des effets globalement non gouvernés lorsqu'il est inséré dans un système d'interaction plus large. La gouvernance de ces propriétés est nécessairement systémique.[4]

La troisième raison est institutionnelle. Les cadres publics émergents ne parient pas sur une autosuffisance comportementale future des modèles. Le framework singapourien de l'IMDA, l'initiative du NIST sur les standards pour les agents, les clarifications du régulateur britannique sur l'application du droit de la consommation aux agents commerciaux, et le cadre juridique européen convergent sur un point : la responsabilité, l'auditabilité, la supervision humaine pertinente et la traçabilité demeurent des exigences du système, non des espoirs placés dans un composant.[5][6][7][8]

Il faut toutefois affiner la thèse initiale. Dire que la gouvernance agentique est exogène ne signifie pas que les propriétés endogènes du modèle sont indifférentes. Elles ne le sont pas. Un système est plus facile à gouverner si ses composants respectent mieux les formats attendus, produisent des justifications plus vérifiables, gèrent plus proprement leur incertitude, ou se conforment davantage à des interfaces structurées. Les progrès endogènes réduisent le coût de la gouvernance ; ils ne la remplacent pas. La relation correcte n'est donc pas celle d'une opposition absolue, mais d'une hiérarchie. La gouvernabilité doit être garantie par l'architecture, même si elle peut être facilitée par les progrès internes des composants.

Le problème n'est donc pas l'agent en soi. C'est le régime d'admission de sa décision dans le système.

3. Action-space, autonomie, réversibilité : une grille de risque pour penser les régimes de décision

Le principal apport conceptuel du cadre de l'IMDA consacré à l'IA agentique est de déplacer l'évaluation du risque depuis le seul modèle vers la combinaison de plusieurs propriétés opérationnelles. Deux d'entre elles sont explicitement centrales : l'étendue de l'action-space et le degré d'autonomie. Une troisième, essentielle en pratique, doit être intégrée dès le départ plutôt qu'ajoutée en note de bas de page : la réversibilité. Le document IMDA définit l'autonomie comme le degré selon lequel un agent peut décider quand et comment agir vers un but, et traite la réversibilité des actions comme une variable structurante de l'analyse de risque.[5]

L'action-space désigne le périmètre des outils, des données, des systèmes et des surfaces d'action auxquels l'agent a effectivement accès. Un agent limité à la consultation en lecture seule d'une base documentaire n'expose pas le même profil de risque qu'un agent capable de modifier un CRM, de reclasser des documents, d'envoyer un message à un client, de valider une transaction ou de publier un contenu dans un environnement externe. L'action-space n'est pas un attribut abstrait. Il définit la portée concrète de ce qui peut se produire en cas de décision erronée, trompeuse, incomplète ou opportuniste.

L'autonomie désigne ici le degré de latitude effectivement laissé au système pour sélectionner ses moyens, enchaîner ses étapes, arbitrer entre options ou déclencher une action sans validation explicite préalable. L'autonomie n'est pas une essence de l'agent. C'est une propriété configurée par le système : niveau d'instruction, permissions, présence de checkpoints, conditions d'escalade, capacité ou non à appeler certains outils, et modalités de validation avant exécution.

La réversibilité qualifie le caractère annulable, corrigeable ou opposable de l'action. Une erreur dans une suggestion de plan de réunion n'a pas le même statut qu'une erreur dans l'envoi d'une notification réglementaire, dans le reclassement massif d'un corpus, dans la modification d'un dossier client, ou dans le déclenchement d'un ordre financier. Certaines actions sont faciles à annuler ; d'autres laissent des traces persistantes, produisent des effets en chaîne, ou créent des conséquences juridiques ou réputationnelles difficilement réparables.

Ces trois variables forment une grille plus robuste que l'opposition simpliste entre "copilot" et "autonomous agent". Elles permettent de raisonner en régimes de décision.

1. Le premier régime est celui de l'assistance. L'agent propose, reformule, synthétise, critique, compare, prépare, mais n'exécute rien au-delà de la production d'un artefact interprétable par l'humain. Son action-space est purement informationnel ou en lecture, son autonomie est faible et ses effets sont hautement réversibles. La gouvernance y est relativement légère, non parce que le système serait intrinsèquement fiable, mais parce que son espace d'impact est fortement borné. Le risque majeur n'est pas l'autonomie technique du système ; c'est la passivité cognitive de l'utilisateur, c'est-à-dire la complaisance ou le rubber-stamping.
2. Le deuxième régime est celui de la recommandation structurée. L'agent ne se contente plus de produire un texte libre ; il pré-remplit un artefact décisionnel, relie ses recommandations à des sources, signale des points d'attention, met en évidence des conflits, prépare un arbitrage ou un diagnostic documentaire. Son action-space peut être plus large, incluant plusieurs corpus, tickets, référentiels, éléments de code ou données métier. Son autonomie est intermédiaire, car il structure le problème et préfigure la décision sans la conclure juridiquement ou opérationnellement. Dans ce régime, la gouvernance exige que chaque recommandation soit rattachable à une provenance, que les règles mobilisées soient identifiables, et que la responsabilité finale demeure clairement humaine. Le risque majeur n'est plus seulement la complaisance ; c'est le biais d'automatisation, autrement dit la tendance à créditer excessivement la proposition du système du simple fait qu'elle est formellement structurée.
3. Le troisième régime est celui de l'exécution bornée. Ici, l'agent agit effectivement sur un environnement, mais dans un périmètre fermé, explicitement autorisé, assorti d'invariants vérifiables et d'un mécanisme d'escalade dès que ces invariants ne sont plus satisfaits. Il peut, par exemple, enrichir des métadonnées selon des taxonomies prédéfinies, déclencher un workflow standard, classer des artefacts documentaires, générer un brouillon formel dans un canevas contraint, ou effectuer une action d'écriture limitée et réversible dans une zone contrôlée. Le cœur de la gouvernance n'est plus alors la seule validation humaine ex ante de chaque action, mais la combinaison d'allowlists strictes, de politiques de

décision déclaratives, d'échantillonnage humain périodique et de journaux d'évidence robustes. Le risque majeur est ici la dérive silencieuse : une erreur faible en apparence mais répétée à grande échelle peut produire une dégradation systémique durable sans incident spectaculaire immédiat.

Ces régimes n'ont pas vocation à constituer une hiérarchie morale ou une trajectoire automatique vers davantage d'autonomie. Ce ne sont pas des niveaux de maturité ; ce sont des modes architecturaux distincts. Un même système peut opérer en assistance sur certaines tâches, en recommandation structurée sur d'autres, et en exécution bornée sur un sous-ensemble très restreint d'actions réversibles. Ce qui importe est que le régime soit explicite, documenté, révisable et justifiable.

Deux modulateurs doivent en outre être ajoutés à cette grille pour en faire un cadre réellement opérationnel.

1. Le premier est la criticité du domaine. Une même configuration action-space, autonomie et réversibilité ne porte pas le même poids selon qu'elle s'applique à du classement interne à faible enjeu, à de la conformité documentaire, à un acte ayant effet sur des droits, ou à un environnement clinique, financier ou légal.
2. Le second est l'asymétrie d'erreur. Certains systèmes supportent relativement bien des faux positifs ou des faux négatifs. D'autres non. Une architecture de gouvernance sérieuse ne peut pas ignorer ces distributions de dommage.

La conséquence est claire. La gouvernance agentique ne doit pas partir de la question "jusqu'où peut-on laisser faire l'agent ?", mais de la question "dans quel régime de décision cet artefact peut-il être admis, à quelles conditions, avec quel périmètre d'action, quelle réversibilité et quel mode d'escalade ?". Ce renversement est la condition de toute gouvernance adulte.

4. De la gouvernance déclarative au contrôle d'admission : vers des contrats de décision

Une fois admis que la gouvernabilité doit être produite par le système, il reste à déterminer quelle forme architecturale peut lui donner une consistance opérationnelle. L'analogie la plus féconde ne se trouve pas dans les métaphores psychologiques de l'agent "responsable" ou "aligné", mais dans certains acquis de l'ingénierie logicielle distribuée. Lorsqu'un environnement technique doit gouverner l'accès de composants à des ressources partagées, il ne présume pas leur vertu. Il institue des mécanismes de contrôle d'admission, d'autorisation, de politique déclarative et d'audit.

La comparaison avec Kubernetes doit être maniée avec précision. Il ne s'agit pas de prétendre qu'un agent et un container sont du même ordre. Un container exécute du code déterministe sur un input donné ; un agent LLM opère dans un espace stochastique et

partiellement interprétatif, où le même input peut produire des séquences d'actions différentes. L'analogie ne porte donc ni sur la cognition ni sur le déterminisme d'exécution, mais sur la structure de gouvernance : dans les deux cas, on ne demande pas à un composant d'être moralement sûr ; on borne ce qu'il peut faire, à quelles conditions et selon quelles politiques déclaratives.

C'est dans cette logique que la notion de contrat de décision prend son sens. Il ne faut pas la présenter comme un standard industriel déjà stabilisé. Ce n'en est pas un. Il s'agit plus rigoureusement d'une proposition de pattern architectural dérivée de la logique du design-by-contract formalisé par Bertrand Meyer, qui transpose l'idée de préconditions, postconditions et invariants du domaine de la correction logicielle vers celui de la gouvernance décisionnelle.[9] Le contrat de décision ne se réduit pas à un log post hoc. Il décrit, préalablement à l'action ou au changement d'état, ce que le système entend faire, sur quelles sources il s'appuie, dans quel régime de décision il opère, sous quelles politiques, avec quels droits, quels invariants, quel niveau de réversibilité et quelles conditions d'escalade. Ce contrat est ensuite soumis à un mécanisme d'admission, déterministe autant que possible, assisté si nécessaire, mais distinct du composant qui demande à agir.

L'intérêt de ce pattern est triple.

1. D'abord, il dissocie fiabilité et gouvernabilité. Un agent peut rester imparfait, variable, même parfois discutable dans ses raisonnements intermédiaires, tout en étant inséré dans un système qui lui interdit d'agir hors du cadre autorisé. La gouvernabilité ne dépend plus du fait que le composant "se comporte bien", mais du fait que le système filtre ce qu'il est admissible de transformer en action.
2. Ensuite, il rend les politiques explicites, versionnées et auditables. Une règle du type "aucune action irréversible ne peut être exécutée en régime d'exécution bornée sans validation humaine nominative" devient un artefact de gouvernance, pas une intention diffuse enfouie dans une documentation ou un prompt. Une politique du type "aucun outil non allowlisté ne peut être appelé depuis ce contexte" devient vérifiable. Le système de gouvernance gagne ainsi en opposabilité.
3. Enfin, il transforme la trace. Un simple log postérieur constate qu'une action a eu lieu. Un contrat de décision bien conçu permet de reconstituer pourquoi une action a été proposée, avec quelles sources, sous quel régime, avec quelle politique applicable, dans quel contexte et avec quelle validation. En environnement régulé, cette différence est décisive. Ce que les organisations doivent être capables de produire n'est pas seulement l'historique brut des actions, mais la chaîne d'admissibilité qui a rendu ces actions possibles.

Il faut toutefois éviter de sur-vendre cette approche. Les contrats de décision, à ce stade, relèvent davantage d'un pattern d'architecture prometteur que d'un standard industriel

mature. Les implémentations existantes sont encore fragmentaires, souvent expérimentales, parfois portées par des projets démonstratifs plus que par des déploiements documentés à grande échelle. Il serait donc abusif d'en parler comme d'un acquis stabilisé de l'industrie. Leur intérêt réside précisément dans leur statut intermédiaire : ils offrent un langage de conception crédible pour transformer des principes de gouvernance en primitives système.

Des expérimentations de petite échelle montrent néanmoins qu'un tel pattern est implémentable. Lorsqu'un agent de veille, de cadrage ou de rédaction n'est pas autorisé à faire progresser seul un artefact vers l'étape suivante sans passage par un contrôle de forme, de rigueur ou de conformité, le système pratique déjà, à une certaine échelle, une gouvernance par admission. Cela ne prouve pas la validité générale du pattern. Mais cela indique qu'il ne relève pas de la pure abstraction.

La vraie difficulté n'est pas conceptuelle. Elle est économique et technique. Introduire une couche de contrôle d'admission ajoute de la latence, de la complexité, un coût de politique, un coût de maintenance et souvent un coût humain. Un système agentique gouverné est plus cher à concevoir, à faire évoluer et à superviser qu'un système agentique laissé à lui-même. Ce surcoût n'est pas un accident regrettable. Il est le prix de la gouvernabilité.

5. Les cadres institutionnels de 2026 : convergence partielle, logiques différentes

L'année 2026 voit émerger plusieurs cadres institutionnels pertinents pour penser la gouvernance des agents, mais ces cadres n'ont ni le même statut, ni la même force normative, ni la même granularité.

Le cadre proposé par l'IMDA à Singapour constitue à ce jour l'une des formulations les plus directement orientées vers les systèmes agentiques. Sa valeur principale n'est pas tant de découvrir ex nihilo des principes inconnus que de les agencer autour de variables opérationnelles adaptées à l'agentique, notamment le périmètre d'action, le degré d'autonomie et la nécessité de mécanismes de contrôle humain significatifs. Son régime est celui de la soft law structurante : il recommande, il cadre, il oriente, mais il ne sanctionne pas par lui-même. Sa force est de proposer une grammaire de conception. Sa limite est de ne pas fournir en tant que tel un régime juridique coercitif.[5]

L'initiative du NIST sur les standards pour agents poursuit une logique différente. Elle ne se situe pas prioritairement sur le terrain de la contrainte juridique, mais sur celui de l'interopérabilité, de la confiance technique, de la sécurité et de la standardisation des interfaces ou mécanismes qui permettront aux systèmes agentiques d'être déployés dans des environnements plus prévisibles. Le NIST la présente explicitement comme une

initiative destinée à favoriser une adoption confiante, un fonctionnement sécurisé et une interopérabilité fluide des agents dans l'écosystème numérique. Il s'agit d'un travail d'infrastructure normative au sens large, plus proche de la fabrique de standards que de la police administrative. Son importance est réelle, mais d'un autre ordre que celle d'une réglementation contraignante.[6]

Le cas de l'Union européenne est encore différent. Le règlement européen sur l'intelligence artificielle introduit un régime juridique contraignant fondé sur une classification par niveau de risque et par catégories d'usage. Il faut ici corriger une simplification fréquente. Un système n'est pas high-risk parce qu'il est "agentique", "autonome" ou "multi-étapes" en tant que tel. Le critère juridique pertinent n'est pas la seule forme architecturale du système, mais son inscription éventuelle dans les catégories ou contextes prévus par le règlement, notamment les cas listés à l'Annexe III, ou son intégration dans certains produits régulés. La Commission européenne rappelle que l'Annexe III recense les cas d'usage high-risk et que ces catégories font l'objet d'un cadrage spécifique.[8]

Il s'ensuit qu'un agent autonome peut ne pas relever du régime high-risk, tandis qu'un système moins spectaculaire mais déployé dans un contexte juridiquement sensible peut y relever pleinement. ***L'agenticité n'est donc pas un critère juridique autonome de qualification***. Cette précision n'affaiblit pas la thèse de l'article. Elle la rend plus robuste. Car même sans assimiler abusivement l'ensemble des agents autonomes à des systèmes high-risk, le cadre européen conforte l'idée qu'un système déployé dans des usages sensibles, ou produisant des effets significatifs sur des droits, obligations ou accès, sera plus susceptible d'entrer dans des régimes exigeant supervision humaine, traçabilité, documentation et accountability.[8]

Le Royaume-Uni, à travers la guidance de la Competition and Markets Authority sur l'usage d'agents IA au regard du droit de la consommation, ajoute une clarification utile. L'agent n'ouvre pas une zone de non-droit. Il constitue un mode d'exécution ou d'intermédiation qui demeure soumis aux obligations existantes du droit de la consommation, et la CMA rappelle explicitement que des manquements peuvent exposer les entreprises à des mesures d'enforcement et à des sanctions.[7]

Pris ensemble, ces cadres ne convergent pas vers un droit mondial unifié de l'agentique. Ils convergent vers une idée plus modeste mais plus importante : aucun acteur institutionnel sérieux ne traite la gouvernance agentique comme une simple question de qualité interne des modèles. Tous la traitent, sous des formes différentes, comme un problème de système.

6. Limites de la thèse et programme de recherche

La thèse défendue ici n'est ni close ni complète. Elle repose sur un ensemble d'inférences plausibles, appuyées par des travaux émergents, des cadres institutionnels et des analogies d'ingénierie, mais elle ne doit pas être présentée comme une démonstration empirique achevée.

1. La première limite est l'absence de données longitudinales robustes sur la relation entre scaling, alignement et fiabilité agentique. Les travaux récents documentent un écart empirique actuel entre capacité et fiabilité sur certains benchmarks et certaines familles de modèles. Ils n'établissent pas une impossibilité théorique de réduction de cet écart à moyen terme. La thèse d'une gouvernance exogène forte doit donc être formulée comme une exigence prudente et architecturale dans l'état présent des systèmes, non comme une réfutation définitive de toute convergence future partielle.[2]
2. La deuxième limite est l'insuffisance de cas industriels documentés. Les patterns proposés ici, notamment la combinaison de régimes de décision, de contrats de décision et de contrôle d'admission, sont intellectuellement cohérents, mais encore faiblement étayés par des publications comparatives montrant leur impact mesuré sur la fréquence des incidents, la charge de conformité, la vitesse opérationnelle ou le coût total de possession. Le passage du prescriptif à l'empirique reste à faire.
3. La troisième limite est économique. Une architecture de gouvernance sérieuse a un coût. Elle ajoute de la friction, de la dette opératoire, des couches de politique, des mécanismes d'escalade, des exigences documentaires, des opérations d'audit et souvent des dépenses humaines non triviales. Tant qu'aucune littérature solide ne compare systématiquement ce coût à celui de la non-gouvernance dans différentes classes de systèmes, l'argument restera partiellement prudentiel.
4. La quatrième limite est liée au domaine de validité du cadre proposé. La grille action-space, autonomie et réversibilité fonctionne bien pour des systèmes enterprise contemporains relativement bornés. Elle sera plus difficile à appliquer à des systèmes auto-modifiants, à des agents traversant plusieurs frontières de confiance organisationnelles, à des environnements multi-agents présentant des comportements émergents difficilement déductibles, ou à des architectures où les permissions elles-mêmes peuvent être dynamiquement reconfigurées par le système. Le cadre n'est pas invalide pour autant ; il est situé.
5. La cinquième limite est géographique et documentaire. Les sources mobilisées dans ce champ restent très largement anglophones, complétées de manière notable par Singapour. Les cadres asiatiques non anglophones sont sous-représentés dans la cartographie courante : les lignes directrices du AI Strategy Council japonais, le cadre réglementaire chinois porté conjointement par le MIIT et la CAC, et les travaux préparatoires coréens en vue d'un AI Act national

constituent autant de sources pertinentes que la présente analyse ne mobilise pas. Cette asymétrie limite la prétention à l'exhaustivité et constitue en elle-même un biais méthodologique à déclarer.

6. La sixième limite porte sur la fragilité de la gouvernance elle-même. L'article postule qu'une architecture de gouvernance exogène bien conçue peut compenser les insuffisances des modèles. Mais cette architecture est elle-même un artefact construit, maintenu et opéré par des organisations qui ne sont pas immunisées contre leurs propres défaillances. Des politiques trop lâches parce que calibrées sous pression opérationnelle, des allowlists qui s'élargissent par accrétion sans revue périodique, des mécanismes d'escalade systématiquement contournés par habitude, des contrats de décision formellement corrects mais substantivement vides, constituent autant de modes de défaillance de la gouvernance par la gouvernance. La qualité d'un système de contrôle d'admission ne dépend pas seulement de ses primitives techniques ; elle dépend de la discipline organisationnelle qui les maintient. Cette récursion, « qui gouverne la gouvernance ? » n'invalide pas l'approche, mais elle interdit de la présenter comme autosuffisante. Une architecture de gouvernance est une condition nécessaire, pas une garantie.

Ces limites n'annulent pas la thèse. Elles définissent plutôt un programme de recherche.

1. La première question est empirique : dans quelle mesure l'écart entre capacité et fiabilité se réduit-il effectivement d'une génération de modèles à l'autre, une fois contrôlés les changements de benchmark, de contexte et de protocole d'évaluation ?
2. La deuxième est architecturale : les patterns de contrôle d'admission, de politique déclarative et de contrat de décision se transposent-ils efficacement au domaine stochastique sans produire une complexité prohibitive ?
3. La troisième est économique : à partir de quel seuil de criticité, de fréquence d'usage et d'asymétrie d'erreur la gouvernance exogène devient-elle rationnelle au regard de son coût total ?
4. La quatrième est systémique : comment gouverner un ensemble multi-agents lorsque le comportement global du système n'est plus inférable à partir des propriétés locales de chacun de ses composants ?

Une discipline de la gouvernance agentique ne naîtra pas d'un seul papier, encore moins d'une seule taxonomie. Elle naîtra de l'articulation entre ingénierie, droit, théorie des systèmes, économie de la conformité et retour d'expérience industriel.

7. Le prix de la gouvernabilité

La conclusion peut maintenant être formulée de manière plus précise.

La bonne question n'est pas : comment rendre les agents suffisamment fiables pour qu'on puisse enfin leur faire confiance sans réserve ? Posée ainsi, la question reconduit une illusion tenace, celle selon laquelle le problème central serait psychologique ou moral, comme si le système devait finir par mériter une confiance quasi personnelle. La question pertinente est plutôt : comment architecturer un système dans lequel aucune décision significative n'accède à l'exécution sans avoir traversé un régime d'admissibilité proportionné à son périmètre d'action, à son niveau d'autonomie, à sa réversibilité, à sa criticité et à l'asymétrie de ses erreurs possibles ?

La confiance, dans cette perspective, n'est pas une propriété héroïque du modèle. Elle est une propriété construite du système. Elle ne suppose pas que le composant soit parfait. Elle suppose que ce qu'il peut transformer en action soit borné, qualifié, journalisé, contrôlé, révisable et attribuable. Elle suppose aussi, dans les systèmes à haut risque, que certaines classes de décision demeurent structurellement indisponibles à l'autonomie pleine et restent conditionnées à une garantie humaine explicite, non comme concession psychologique à la prudence, mais comme principe d'organisation de la responsabilité.

Cette thèse n'est pas confortable. Elle renonce à l'une des promesses les plus séduisantes de l'imaginaire agentique contemporain, celle d'une autonomie croissante qui éliminerait progressivement le goulot humain. Ce renoncement est pourtant moins un conservatisme qu'une clarification. Dans les systèmes à conséquences élevées, le goulot humain n'est pas un défaut transitoire appelé à disparaître avec l'amélioration des modèles. Il constitue une garantie institutionnelle. Dans les domaines où une décision peut engager la sécurité collective, l'intégrité d'une infrastructure critique, la conduite d'une opération militaire, l'exposition nucléaire, ou l'état de santé d'un patient, aucune architecture sérieuse ne devrait autoriser un système à prendre seul la décision finale sans supervision humaine significative. Ce qui peut devenir autonome, sous conditions, n'est donc pas la décision souveraine au sens plein. C'est l'exécution bornée de contraintes, de séquences ou d'opérations déjà légitimées, sous un régime de contrôle où l'intervention humaine demeure le verrou ultime de l'admissibilité.

Le coût de cette gouvernabilité est réel. Complexité architecturale, latence, maintenance de politiques, validation humaine, discipline documentaire, friction organisationnelle. Ce coût n'est pas un défaut secondaire. Il est le prix exact d'un système capable de rendre compte de ce qu'il a fait, pourquoi il l'a fait, dans quel cadre il l'a fait, et qui en assumait la responsabilité.

Les modèles continueront à progresser. Il serait absurde de parier contre cela. Mais même des modèles meilleurs ne supprimeront pas la nécessité d'une architecture de gouvernance. Au mieux, ils en réduiront le coût marginal.

Un agent n'est pas gouverné parce qu'il est meilleur. Il est gouverné parce qu'un système lui interdit d'agir hors d'un régime décisionnel explicitement autorisé.

Dans les systèmes à haut risque, l'autonomie ne doit pas supprimer le goulot humain. Elle doit apprendre à vivre sous lui.

Notes

[1] Gartner, "Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5% in 2025," press release, 26 août 2025. https://www.gartner.com/en/newsroom/press-releases/2025-08-26-gartner-predicts-40-percent-of-enterprise-apps-will-feature-task-specific-ai-agents-by-2026

[2] Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan, "Towards a Science of AI Agent Reliability," arXiv:2602.16666, 18 février 2026. https://arxiv.org/abs/2602.16666

[3] Simon Willison, "Writing about Agentic Engineering Patterns," 23 février 2026 ; voir aussi *Agentic Engineering Patterns*, guide en cours, simonwillison.net. Références mobilisées comme observations de praticien, non comme mesure de marché. https://simonwillison.net/guides/agentic-engineering-patterns/

[4] Gartner présente AI TRiSM comme un cadre général destiné à assurer gouvernance, trustworthiness, fairness, fiabilité, robustesse, efficacité et protection des données dans les déploiements d'IA. La présente discussion en prolonge la logique au cas des systèmes agentiques dans un sens doctrinal et systémique ; il ne s'agit pas d'attribuer à Gartner une formalisation agentique détaillée qu'il ne revendique pas explicitement dans cette source. Voir Gartner, "Tackling Trust, Risk and Security in AI Models," 24 décembre 2024.

[5] Infocomm Media Development Authority, *Model AI Governance Framework for Agentic AI*, version 1.0, 22 janvier 2026, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf. Le framework est également référencé sur la page institutionnelle d'IMDA consacrée à l'intelligence artificielle.

[6] National Institute of Standards and Technology, Center for AI Standards and Innovation, "Announcing the AI Agent Standards Initiative: Interoperable and Secure," 17 février 2026. https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure

[7] UK Competition and Markets Authority, "Complying with Consumer Law When Using AI Agents," 9 mars 2026. https://www.gov.uk/government/publications/complying-with-consumer-law-when-using-ai-agents

[8] Commission européenne, "Navigating the AI Act," FAQ relative au règlement (UE) 2024/1689, notamment sur les systèmes high-risk et l'Annexe III. https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act

[9] Bertrand Meyer, "Applying Design by Contract," *Computer* 25, no. 10 (1992): 40–51. https://doi.org/10.1109/2.161279

10] Les évaluations publiées par les laboratoires montrent des gains inter-générationnels sur certaines dimensions telles que le suivi d'instructions, la structuration des sorties et la gestion de contraintes de format. Voir par exemple Anthropic, *System Card: Claude Opus 4 & Claude Sonnet 4* (mai 2025, mise à jour juillet 2025), ainsi que OpenAI, *GPT-4o System Card* (8 août 2024). Ces progrès ne doivent toutefois pas être confondus avec des gains équivalents en consistance ou en fiabilité agentique au sens de [2], ni avec la possibilité de supprimer les garanties humaines dans les systèmes à conséquences élevées. Dans les environnements à haut risque, les progrès endogènes des modèles peuvent réduire la charge de supervision, mais ils ne sauraient abolir l'exigence d'un contrôle humain significatif lorsque les décisions engagent la sécurité, l'intégrité d'infrastructures critiques, ou des effets substantiels sur des personnes.

Bibliographie

Anthropic. "Model system cards." Page d'index des system cards, consultée en 2026. https://www.anthropic.com/system-cards.

Anthropic. *System Card: Claude Opus 4 & Claude Sonnet 4*. Mai 2025, mise à jour 16 juillet 2025. https://www.anthropic.com/claude-4-system-card.

Commission européenne. "Navigating the AI Act." FAQ relative au règlement (UE) 2024/1689. https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act.

Competition and Markets Authority. *Complying with Consumer Law When Using AI Agents*. GOV.UK, 9 mars 2026. https://www.gov.uk/government/publications/complying-with-consumer-law-when-using-ai-agents.

Gartner. "Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5% in 2025." Press release, 26 août 2025. https://www.gartner.com/en/newsroom/press-releases/2025-08-26-gartner-predicts-40-percent-of-enterprise-apps-will-feature-task-specific-ai-agents-by-2026-up-from-less-than-5-percent-in-2025.

Gartner. "Tackling Trust, Risk and Security in AI Models." 24 décembre 2024. https://www.gartner.com/en/articles/ai-trust-and-ai-risk.

Infocomm Media Development Authority. *Model AI Governance Framework for Agentic AI*. Version 1.0. Singapour, 22 janvier 2026. https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf.

Meyer, Bertrand. "Applying Design by Contract." *Computer* 25, no. 10 (1992): 40–51. https://doi.org/10.1109/2.161279.

National Institute of Standards and Technology, Center for AI Standards and Innovation. “Announcing the AI Agent Standards Initiative: Interoperable and Secure.” 17 février 2026. https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure.

OpenAI. “GPT-4o System Card.” 8 août 2024. https://openai.com/index/gpt-4o-system-card/. Version PDF : https://cdn.openai.com/gpt-4o-system-card.pdf.

Rabanser, Stephan, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. "Towards a Science of AI Agent Reliability." arXiv:2602.16666, 18 février 2026. https://arxiv.org/abs/2602.16666.

Willison, Simon. "Writing about Agentic Engineering Patterns." 23 février 2026. https://simonwillison.net/2026/Feb/23/agentic-engineering-patterns/.

Willison, Simon. *Agentic Engineering Patterns*. Guide en cours, 2026. https://simonwillison.net/guides/agentic-engineering-patterns/.