Ce qu'un agent refuse dit plus que ce qu'il fait

La taxonomie de refus comme primitive architecturale de la gouvernance agentique

1. Une métrique presque absente du paysage d'observabilité

L'industrialisation des agents d'entreprise, en 2026, a produit une infrastructure d'observabilité impressionnante, mais structurellement dissymétrique. Les agents déployés dans les grands environnements commerciaux comme dans la plupart des implémentations internes dérivées exposent désormais des métriques fines de capacité : taux de complétion, succès des appels d'outils, latence d'exécution, temps moyen de résolution, fréquence de réessai, coût unitaire, consommation « tokenique », proportion d'escalades, productivité apparente par tâche ou par séquence. Les tableaux de bord qui en résultent donnent à voir l'agent comme une trajectoire instrumentée de performance.

Ce que ces infrastructures exposent beaucoup moins, dans les outils dominants du marché comme dans les pratiques usuelles de déploiement, relève pourtant d'une autre famille de signaux : la distribution de leurs refus, la typologie de ces refus, les conditions qui les déclenchent, la distinction entre un refus propre au système et un blocage imposé par une couche externe, la relation entre refus observés et politique décisionnelle spécifiée, la part des non-exécutions qui relèvent d'une incapacité, d'une abstention légitime, d'une contrainte d'autorité, d'un défaut de contrat ou d'une invalidité temporelle des prémisses. Autrement dit, les systèmes contemporains instrumentent correctement ce que l'agent accomplit ; ils instrumentent beaucoup moins ce à quoi il renonce légitimement.

Cette dissymétrie n'est pas un simple retard de maturité. Elle est cohérente avec l'économie politique du secteur. Un agent se vend par ses succès. Il se montre à travers ses tâches complétées, ses gains de temps, son taux d'automatisation, sa fluidité d'usage. À l'inverse, un agent qui expose massivement ses refus donne facilement l'impression de mettre en scène ses limites. Le marché récompense donc la visibilité de la capacité bien davantage que celle de la non-action légitime. Ce biais commercial n'est pas anecdotique. Il se retrouve de manière convergente dans les tableaux de bord exposés par les principaux frameworks et plateformes d'orchestration agentique, qui privilégient quasi systématiquement les métriques de complétion sur les métriques de non-exécution.

Jérôme Vetillard / Twingital Institute, avril 2026

Le coût de cette dissymétrie apparaît dès que l'on quitte les environnements de productivité à faible enjeu pour entrer dans des contextes à effets asymétriques. Dans ces contextes, un système ne peut être dit gouvernable que si ses conditions d'action et de non-action sont spécifiées, observables, auditables et comparables à une politique explicite. Cette définition mérite d'être posée d'emblée. Elle évite de traiter la gouvernabilité comme un simple mot d'ordre. Un système est gouvernable non parce qu'il est seulement surveillé de l'extérieur, mais parce que sa logique décisionnelle est suffisamment structurée pour que l'organisation puisse savoir quand il agit, quand il s'abstient, pourquoi, et en vertu de quelle clause.

Sous cet angle, une observabilité centrée sur le seul succès est insuffisante. Elle permet de suivre une performance ; elle ne permet pas d'évaluer une discipline décisionnelle. Un système qui maximise ses complétions sans exposer la structure de ses refus peut certes être encadré par des permissions, du sandboxing, des quotas, des validations humaines ou des garde-fous applicatifs. Mais sa logique de décision propre demeure opaque. Sa gouvernance reste alors extrinsèque, partielle, coûteuse à maintenir, et souvent dépendante des incidents pour révéler ses défauts. Elle ne disparaît pas totalement ; elle demeure simplement incomplète à l'endroit même où se joue l'autonomie relative du système.

La thèse de cet article peut alors être formulée avec précision : dans les systèmes agentiques déployés en environnement régulé ou à réversibilité limitée, la maturité de gouvernance se mesure moins par le taux de succès que par la granularité, la qualité d'instrumentation et l'adéquation contractuelle de leurs refus. ***Le refus n'est pas un résidu du système : Il est une sortie décisionnelle à part entière***. Un système qui n'est pas capable d'exposer cette sortie sous une forme typée et auditée n'est pas simplement sous-observé ; il demeure gouvernable principalement par contraintes externes, sans lisibilité suffisante de sa logique décisionnelle interne.

Le domaine de validité de cette thèse doit être explicité sans ambiguïté. Elle concerne les agents opérant dans des environnements régulés ou à effets asymétriques : santé, finance, assurance, industrie critique, administration publique, supply chain sensible, sécurité, ou plus généralement toute opération où le coût d'une action incorrecte excède largement celui d'une non-action temporaire. Elle vaut également dans les contextes à réversibilité limitée, c'est-à-dire là où une action, une fois exécutée, ne peut être annulée qu'au prix d'un coût matériel, juridique, réputationnel ou clinique significatif. Elle vaut beaucoup moins pour les agents purement informatifs, les assistants de productivité à faible enjeu ou les agents conversationnels grand public. Cette restriction n'est pas une prudence rhétorique. Elle est la condition même de la rigueur.

2. Refus, échec, modération, blocage : quatre phénomènes distincts

La littérature industrielle sur les agents emploie souvent de manière interchangeable des termes qui ne désignent pourtant pas la même réalité : refus, échec, garde-fou, abstention, modération, blocage. Cette indistinction affaiblit considérablement toute discussion sérieuse sur la gouvernance, parce qu'elle mélange des phénomènes qui n'occupent pas le même niveau logique.

- **Un succès** est la complétion d'une tâche au regard de ses critères d'acceptation. Ces critères peuvent être fonctionnels, lorsque l'action produit l'effet attendu ; formels, lorsque le format de sortie est conforme ; contractuels, lorsque le résultat reste dans les bornes de ce qui a été demandé et permis ; ou organisationnels, lorsqu'il respecte la politique de délégation applicable. Le succès n'est donc pas seulement un effet obtenu. C'est un effet obtenu dans des conditions recevables.
- **Un échec** est une non-exécution non anticipée par le système. L'agent tente d'agir, ou initie une chaîne d'action, et le résultat attendu n'aboutit pas. L'appel d'outil échoue, le workflow se brise, la donnée requise n'est pas accessible, l'exécution devient incohérente, l'agent entre en boucle, ou produit un résultat non conforme sans avoir identifié ex ante qu'il n'aurait pas dû tenter l'action. L'échec est un événement diagnostiqué après coup.
- **Un refus**, au sens strict retenu dans cet article, est une non-exécution anticipée par le système lui-même, produite avant l'exécution de l'action en vertu d'une condition structurelle reconnue comme bloquante. Le système ne tente pas et n'échoue pas. Il reconnaît qu'il ne doit pas tenter. Le refus est donc un acte décisionnel de non-engagement. Il n'est pas une simple absence d'action ; il est une sortie structurée du système.
- **Une modération *post-hoc***, enfin, est une intervention défensive exercée sur une production déjà engagée ou déjà formulée. Elle opère au niveau d'un filtre de contenu, d'un middleware de sécurité, d'un rail externe, d'un contrôle applicatif, d'un proxy de politique ou d'une couche de supervision qui invalide une sortie ou intercepte un appel. La modération peut être indispensable. Elle n'est pas, pour autant, un refus propre au système. Elle exprime une décision de l'environnement sur le système, et non une décision du système sur sa propre action.

Cette distinction doit être tenue fermement, car elle est fondatrice :

- Le refus structurel vient du système,
- La modération vient du garde-fou,
- L'échec vient d'une exécution qui n'a pas abouti,
- Le succès vient d'une exécution qui a abouti dans des conditions recevables.

Confondre ces quatre catégories conduit à attribuer à l'agent une discipline qu'il n'a peut-être pas, ou à attribuer à l'infrastructure de sécurité une compétence décisionnelle qui ne lui appartient pas. Plus grave encore, cela rend le débat de gouvernance pratiquement inintelligible : on ne sait plus si l'on parle d'un système qui sait s'abstenir, d'un système qui agit puis est bloqué, ou d'un système qui tente et casse.

La précision conceptuelle n'est pas un luxe terminologique. Elle conditionne le modèle de données de l'observabilité. Si ces phénomènes ne sont pas distingués dès la conception, ils se retrouveront fusionnés dans les logs sous des libellés composites du type declined, blocked, failed ou non-completion. À partir de là, toute analyse ultérieure devient partiellement aveugle.

Dans ce texte, le mot refus désignera exclusivement le refus structurel. La modération post-hoc reste nécessaire dans une stratégie de défense en profondeur, mais elle appartient à un autre niveau d'analyse. Elle ne constitue pas la preuve que l'agent sait ne pas faire. Elle constitue seulement la preuve que son environnement sait encore l'arrêter.

3. Pourquoi le taux de succès devient un proxy trompeur

Dès qu'une métrique est simple à calculer, consolidable à grande échelle et politiquement valorisante, elle tend à devenir la fonction objective effective du système qui la regarde. Le taux de succès remplit parfaitement ces conditions. Il s'extrait facilement des traces d'exécution, se présente en pourcentage, s'agrège par équipe, use case, période ou client, et s'insère naturellement dans un récit de progrès. Une hausse de quelques points suffit à produire une narration de maturité, même lorsqu'on ignore ce qui a été sacrifié pour l'obtenir.

Le problème n'est pas que le taux de succès serait inutile mais qu'il devient, dès qu'il occupe le centre du tableau de bord, ***un proxy trop pauvre pour résumer la qualité décisionnelle d'un agent***. Plus précisément, il ne permet pas de distinguer entre deux trajectoires radicalement différentes : celle d'un agent qui réussit davantage parce qu'il est réellement meilleur, et celle ***d'un agent qui réussit davantage parce qu'il a progressivement cessé de reconnaître les situations où il aurait dû s'abstenir***.

C'est ici qu'intervient une forme particulière de la loi de Goodhart : Lorsqu'une mesure devient une cible, elle cesse d'être une bonne mesure (que nous avions déjà citée ici dans un autre contexte). Ce mécanisme, bien documenté dans d'autres systèmes d'optimisation, se transpose ici sous une forme spécifique liée à la nature actionnelle des agents : toute pression continue à la hausse sur les complétions pousse le système, son prompt, son orchestration, son cadre d'évaluation, parfois même son fine-tuning, vers la suppression des comportements de non-engagement. Le refus devient une perte

apparente. Il cesse d'être lu comme une compétence de discipline et commence à être traité comme une friction à éliminer.

Ce phénomène n'est pas sans parenté avec des pathologies déjà décrites dans la littérature sur les modèles génératifs : tendance à répondre plutôt qu'à reconnaître une incertitude, à satisfaire l'évaluateur plutôt qu'à exprimer l'état épistémique réel du système, à produire un contenu recevable en apparence plutôt qu'un signal fidèle de ses limites. Dans le cas d'un assistant textuel, cette dérive conduit à produire une réponse au lieu de reconnaître une ignorance. Dans le cas d'un agent, elle peut conduire à produire une action au lieu de reconnaître qu'aucune action ne devrait être engagée. Le changement n'est pas de degré. Il est de régime. L'artefact ne se contente plus d'énoncer. Il initie.

La conséquence opérationnelle est nette : la courbe de progression du taux de succès, prise isolément, est un indicateur ambigu. Elle peut témoigner d'un progrès réel de capacité. Elle peut aussi masquer l'érosion d'une compétence plus discrète mais décisive : la compétence de refus. Tant que la distribution des refus n'est pas instrumentée parallèlement, il est impossible de départager ces deux interprétations.

Une objection classique consiste à dire que les systèmes ne mesurent pas seulement leurs succès, mais aussi leurs erreurs. L'objection est juste, mais elle ne répond pas au point central. Une erreur est un événement subi ou constaté après tentative. Un refus est une décision de non-engagement produite avant tentative. Mesurer les erreurs revient à observer ce que le système n'a pas su empêcher. Mesurer les refus revient à observer ce qu'il a su s'interdire. Cette différence n'est pas quantitative. Elle est ontologique. Et c'est précisément cette différence qui sépare un système piloté par incidents d'un système partiellement gouvernable par ses sorties décisionnelles.

4. Ce que les infrastructures actuelles exposent mal

L'asymétrie diagnostiquée plus haut n'est pas seulement une impression générale. Elle peut être caractérisée plus précisément à travers quatre zones typiques d'invisibilité.

1. La première est l'invisibilité de la distribution des refus par mécanisme. Lorsqu'un agent n'accomplit pas une tâche, l'observateur dispose rarement d'un moyen fiable pour distinguer entre un refus par domaine d'applicabilité, une absence d'autorité, un blocage par politique externe, une permission insuffisante sur un outil, un fallback silencieux, un timeout, une dégradation capacitaire ou un simple échec d'exécution. Tous ces phénomènes tendent à se retrouver agrégés dans une même famille de non-completion. Cette fusion détruit immédiatement la valeur diagnostique de la trace.

2. La deuxième est l'absence de contexte décisionnel explicite pour les non-exécutions. Un refus digne de ce nom ne devrait pas apparaître comme un événement nu. Il devrait être accompagné, au minimum, du type de refus émis, du signal décisionnel mobilisé, de la borne contractuelle pertinente, de l'horodatage des prémisses, du contrat de décision invoqué, et du canal d'escalade éventuellement déclenché. Dans les pratiques actuelles, la non-exécution est très souvent visible comme résultat, beaucoup plus rarement comme décision documentée.
3. La troisième est l'amalgame logique entre refus, modération et blocage dans les modèles de journalisation. Beaucoup d'outils d'observabilité héritent d'une vision du système agentique comme simple appel d'API enrichi. Dans ce cadre, la notion de non-réponse ou de réponse bloquée suffit. Mais un agent n'est pas seulement un générateur de texte outillé. C'est un système qui arbitre entre plusieurs registres de sortie : agir, demander, escalader, attendre, suspendre, ne pas faire. Le modèle de données doit refléter cette pluralité. Sans cela, l'observabilité reste au niveau LLM, alors que le problème de gouvernance se situe au niveau décisionnel.
4. La quatrième est la non-exposition de la politique décisionnelle elle-même. Les conditions sous lesquelles un système s'abstient sont aujourd'hui dispersées entre les poids du modèle, le prompt système, les règles du framework, les permissions applicatives, les configurations des outils, les middlewares de sécurité et parfois la logique métier externe. Il en résulte que le déployeur hérite d'une politique de non-exécution qu'il ne peut pas lire comme un artefact unifié. Il l'infère après coup, empiriquement, au gré des cas rencontrés. Cette situation est acceptable pour un chatbot grand public. Elle l'est beaucoup moins pour un système appelé à produire des effets dans un environnement régulé.

Ces quatre zones convergent vers une conclusion simple. L'infrastructure actuelle expose principalement la capacité, beaucoup moins la gouvernabilité. En contexte sensible, le déploiement d'un agent ne peut donc pas reposer sur les sorties natives du produit. Il exige une couche de structuration supplémentaire. Cette couche ne peut être conçue sérieusement qu'à partir d'une taxonomie de refus explicite.

5. Refus épistémiques, normatifs, pragmatiques : le besoin d'une taxonomie structurale

Une taxonomie utile doit satisfaire trois conditions :

- Elle doit être suffisamment exhaustive pour couvrir les situations effectivement rencontrées dans les contextes visés,

- Elle doit être suffisamment exclusive pour permettre un triage relativement univoque des cas dominants, même si certains événements cumulent plusieurs motifs,
- Enfin, elle doit classer des mécanismes, non des contenus. Une taxonomie qui distingue des refus « médicaux », « juridiques » ou « financiers » ne classe pas des propriétés du système. Elle classe des secteurs. Or la gouvernance architecturale a besoin d'une typologie structurale.

Avant de présenter les catégories, une distinction intermédiaire est utile. Tous les refus ne relèvent pas du même registre.

- Certains sont épistémiques : le système s'abstient parce qu'il ne dispose pas de conditions cognitives suffisantes pour agir de manière recevable,
- D'autres sont normatifs ou contractuels : le système s'abstient parce qu'aucune clause de décision ne lui donne légitimité pour agir,
- D'autres encore sont organisationnels ou d'autorité : le système s'abstient parce que l'organisation n'a pas délégué le niveau requis,
- Enfin, certains sont pragmatiques ou techno-opérationnels : le système s'abstient parce que l'action n'est plus valable dans le temps utile ou ne peut plus être réversible dans l'enveloppe admise.

Cette stratification évite de faire croire que tous les refus seraient homogènes. Ils ne le sont pas. Mais ils peuvent être regroupés en six catégories fonctionnelles couvrant l'essentiel de l'espace opérationnel.

5.1 Refus par domaine d'applicabilité

Le système rencontre une entrée, une situation ou une configuration qu'il n'a pas de raison suffisante de considérer comme appartenant au domaine sur lequel sa politique de décision a été validée. Le mécanisme peut mobiliser une mesure de distance, de densité, d'incertitude, de dérive, d'appartenance contextuelle, ou un autre indicateur d'écart à l'espace de validité. Ce qui importe ici n'est pas la technique précise, mais le fait qu'une clause explicite relie cet écart à une non-exécution.

La signature observable d'un tel refus doit comporter non seulement le type *out-of-domain*, mais également le signal mobilisé, sa valeur, la borne pertinente et le contrat qui lui donne son statut. Sans cela, le refus reste une intuition encapsulée dans le système plutôt qu'une sortie auditable.

5.2 Refus par réversibilité insuffisante

L'action demandée est peut-être faisable, mais elle excède l'enveloppe de réversibilité accordée à l'agent. La question n'est pas seulement de savoir si l'action est, en théorie, réversible. Elle est de savoir si elle l'est dans la fenêtre temporelle, le coût de rollback, la

portée d'effet et le cadre de responsabilité prévus. Déplacer un fichier peut être autorisé. Vider irréversiblement une corbeille ou purger un enregistrement clinique ne l'est pas nécessairement. Annuler une commande dans les deux heures peut être admis. Pas au-delà.

Ce refus traduit, en termes architecturaux, un principe de précaution opérationnelle. Il est particulièrement important dans les systèmes où la délégation ne doit pas être définie seulement par la nature de l'action, mais par sa récupérabilité effective.

Cela aurait pu éviter certains gros incidents de production chez des Hyperscalers dont la production est de plus en plus pilotée par des agents IA mais non dotés de ce type de critères issus de leur architecture de gouvernance.

5.3 Refus par latence décisionnelle excédée

Une décision n'est pas seulement liée à un contenu. Elle est liée à un moment. Il existe des situations où agir à partir de prémisses devenues trop anciennes revient à agir sans base valide. La non-exécution correcte ne provient alors ni d'un manque d'autorité ni d'un manque de capacité, mais d'une invalidité temporelle des fondements de la décision.

Cette catégorie est souvent sous-estimée, alors qu'elle est décisive dans de nombreux environnements : flux de prix, constantes vitales, états logistiques, données de supervision industrielle, conditions de marché, disponibilité de stock, validité documentaire. La temporalité n'est pas une métadonnée secondaire ; elle est une propriété de validité de la décision.

5.4 Refus par signal décisionnel sous seuil contractuel

Cette catégorie exige une précision technique. Ce n'est pas le refus lui-même qui est calibré au sens strict. Ce sont les signaux sur lesquels repose la décision, lorsque ces signaux peuvent être rendus interprétables, c'est-à-dire lorsque leur valeur peut être reliée de manière stable à une propriété empirique (fréquence, erreur, conformité, robustesse). L'agent ne « calibre » pas son refus ; il émet un refus parce qu'un signal décisionnel interprétable, éventuellement calibré, se situe sous le seuil requis pour la classe d'action considérée.

La distinction n'est pas cosmétique. Dans les cas où une calibration probabiliste est possible et pertinente, elle doit être réalisée en amont sur les scores utilisés pour décider. Le refus survient ensuite comme conséquence contractuelle de la comparaison entre ce signal et une borne spécifiée. Dans d'autres cas, le signal pourra relever non d'une probabilité calibrée mais d'un score de conformité, de similarité, de robustesse ou de validité contextuelle, pourvu que sa sémantique et son usage soient documentés.

5.5 Refus par absence de contrat de décision applicable

Le système se trouve dans une situation pour laquelle aucune clause de décision n'existe. Ce refus est plus fondamental qu'il n'y paraît. Il ne signale pas seulement une lacune de capacité. Il signale une lacune de gouvernance. Le système rencontre un cas que l'organisation n'a pas explicitement couvert. La bonne réponse n'est donc pas l'improvisation, mais la non-exécution documentée.

Cette catégorie joue un rôle particulier, car elle constitue un méta-refus. Elle ne dit pas seulement : « je ne peux pas agir dans ce cas ». Elle dit : « aucune politique légitime ne me permet ici d'arbitrer ». En ce sens, elle protège le système contre la tentation de combler par initiative locale un vide normatif qui devrait être traité au niveau de la conception.

5.6 Refus par autorité insuffisante

Certaines actions relèvent d'une délégation graduée. Le système peut être techniquement capable, cognitivement assez sûr, temporellement valide, et pourtant non autorisé à agir seul. La non-exécution correcte vient alors du fait que le niveau d'autorité requis n'a pas été accordé. Ce refus ne constitue pas un simple fallback. Il matérialise une clause explicite de répartition des rôles entre système et humain, ou entre plusieurs niveaux d'autorité.

Ce type de refus est crucial, car il rend visible le point précis où la gouvernance humaine verrouille l'autonomie du système. Dans les domaines à haut risque, cette clause ne doit pas être traitée comme une humiliation de l'agent, mais comme l'institution d'un goulot de responsabilité.

Ces six catégories ne prétendent pas à l'orthogonalité parfaite. Une même situation peut déclencher plusieurs motifs. Un cas peut être simultanément hors domaine, sous seuil sur le signal de décision et au-dessus du niveau d'autorité délégué. Mais elles sont distinctes par leur structure, par leurs conditions de déclenchement et par les artefacts d'audit qu'elles supposent. C'est cette observabilité différentielle qui les rend utiles.

5bis. Un terrain d'implémentation : l'applicabilité multi-signal de ToxTwin

La taxonomie proposée resterait purement conceptuelle si aucune de ses catégories ne disposait d'une traduction opérationnelle. Or il existe au moins une classe de systèmes dans laquelle une partie de cette structure a déjà été instanciée, sous une forme certes partielle mais suffisante pour établir une preuve de faisabilité : les scoreurs de toxicité en chimioinformatique, et plus précisément, dans le travail de R&D mené autour de ToxTwin, un dispositif d'applicabilité multi-signal destiné à décider, pour une molécule candidate, si le système a le droit de produire une prédiction de toxicité ou s'il doit s'abstenir.

Le dispositif combine trois signaux indépendants. Une similarité chimique de Tanimoto calculée sur les empreintes structurales mesure à quel point la molécule interrogée ressemble à ce que le système a déjà rencontré. Une distance dans l'espace latent d'un réseau de neurones sur graphes, agrégée par k plus proches voisins, mesure la proximité de la molécule aux représentations apprises. Une estimation de densité dans cet espace mesure la rareté relative de la région rencontrée. Les trois signaux sont combinés selon une règle explicite qui produit un verdict typé : dans le domaine, en marge du domaine, ou hors domaine. Seuls les cas « dans le domaine » autorisent la production d'une prédiction ; les cas marginaux déclenchent un signalement ; les cas hors domaine produisent un refus typé, journalisé, accompagné des valeurs des trois signaux et de la règle qui les compose.

La vérification opérationnelle la plus parlante est venue du comportement du système sur les complexes de coordination métalliques, dont les trois molécules de chimiothérapie platinée utilisées en oncologie offrent un cas d'école : cisplatine, carboplatine, oxaliplatine. Ces composés appartiennent à une classe chimique absente du domaine d'apprentissage du modèle. La featurisation utilisée, dérivée de standards pensés pour des molécules organiques, ne représente pas correctement les liaisons de coordination du platine. Le dispositif d'applicabilité a rangé les trois molécules hors domaine, exactement comme il était structurellement censé le faire. Le système ne prétend pas connaître la toxicité de ces molécules ; il signale qu'il n'est pas légitime à la prédire dans cette configuration. C'est précisément cette réponse qui est utile.

Ce que cet exemple montre est circonscrit. Il montre qu'une catégorie de la taxonomie, le refus par domaine d'applicabilité, est implémentable avec des briques techniques actuelles, journalisable sous une forme typée, et opérationnellement robuste sur au moins une famille de cas hors distribution. Il ne montre pas que l'ensemble de la taxonomie est déjà couvert, que la calibration inter-domaines des seuils est résolue, ni que cette preuve d'existence se transpose sans coût à d'autres systèmes agentiques. La contribution de l'exemple n'est pas de généraliser. Elle est de montrer que l'écart entre la thèse architecturale et son implémentation n'est pas infranchissable.

6. Le refus comme sortie de même rang logique que l'action

Le refus n'est pas le dehors de la décision. Il en est l'un des modes.

Le point décisif n'est pas seulement qu'il existerait différents types de refus. Le point décisif est que le refus doit être traité comme une sortie décisionnelle de même rang logique que l'action. Tant que le refus reste conçu comme une absence, un fallback, une

panne douce ou un résidu de prudence, il demeure architecturalement subalterne. Il n'est ni contractualisé, ni vérifiable, ni comparable à une politique explicite.

Il faut donc reformuler le contrat de décision en conséquence.

Par contrat de décision, nous entendons ici un artefact formalisable, potentiellement implémentable sous une forme de « policy-as-code », c'est-à-dire une politique exprimée sous forme exécutable et versionnable, qui relie une classe de situations à un ensemble de registres de sortie légitimes, à leurs conditions de déclenchement, à leurs obligations de journalisation et à leurs règles d'escalade. Le contrat ne décrit pas seulement ce que l'agent peut faire ; il décrit aussi ce qu'il doit ne pas faire, dans quelles conditions, sous quelle forme, et vers quel canal.

Un contrat de décision valide ne doit donc pas seulement spécifier ce que l'agent est autorisé à faire lorsque certaines préconditions sont satisfaites. Il doit également spécifier les conditions sous lesquelles l'agent doit ne pas faire, le type de refus qu'il doit alors produire, les signaux ou bornes qui le déclenchent, les métadonnées qu'il doit journaliser, le canal vers lequel il doit escalader et, le cas échéant, la temporalité au terme de laquelle la situation devra être réévaluée.

Autrement dit, le contrat de décision ne relie pas seulement un contexte à une action. Il relie un contexte à un espace fini de sorties légitimes : action, refus typé, escalade, suspension, demande d'information complémentaire, voire transmission humaine lorsqu'une politique l'exige. Le refus devient ainsi l'un des modes explicitement admis de la décision, non son dehors.

Cette reformulation produit plusieurs conséquences structurantes. La première est l'auditabilité par conception. Si un contrat stipule qu'en dessous d'un certain seuil sur un signal décisionnel interprétable, la classe d'action A doit produire un refus de type déterminé, alors l'absence de ce refus dans un cas où le signal était sous la borne devient elle-même un incident objectivable. Ce qui se dissout aujourd'hui dans la masse des décisions prises « quand même » peut demain devenir une non-conformité décisionnelle.

La deuxième est la vérifiabilité de couverture. On peut demander, pour chaque classe de décision, quels types de refus ont été explicitement adressés, lesquels ont été exclus, et pour quelles raisons. Cette propriété est centrale. Elle permet de distinguer une zone de silence pensée d'une zone de silence subie.

La troisième est la mesurabilité de l'écart entre politique et comportement observé. Une fois les refus contractuellement définis, leur distribution empirique peut être comparée à la distribution attendue. Les écarts peuvent alors signaler des dérives : dérive de domaine, dérive d'usage, dérive d'orchestration, régression du système, dégradation d'un signal, inadéquation progressive du contrat au réel rencontré.

Il faut ici répondre à une objection prévisible. Cette formalisation des refus ne risque-t-elle pas de rigidifier les agents au point de les rendre peu utiles ? La réponse est non, à condition de distinguer rigidité de comportement et rigidité de structure. Ce qui est rigidifié n'est pas le détail des séquences internes de l'agent. Ce qui est rigidifié, c'est la forme légitime de ses sorties. L'agent demeure libre, à l'intérieur de ses bornes, d'explorer, de planifier, d'appeler des outils, de reconfigurer ses sous-tâches, de reformuler une demande ou de choisir un parcours d'exécution. Ce qui est fixé, c'est le registre de décision auquel il est autorisé à aboutir, et les conditions sous lesquelles un registre plutôt qu'un autre devient recevable. Cette rigidité-là n'est pas une faiblesse. Elle est la condition de possibilité du déploiement dans les environnements où l'autonomie ne peut jamais être laissée sans lisibilité.

7. Le refus comme primitive d'architecture, et non comme effet secondaire du modèle

La contribution de cette thèse doit être située correctement. Elle ne consiste pas à « découvrir » le refus. Les traditions de l'apprentissage statistique connaissent depuis longtemps des formes d'abstention sélective, d'option de rejet, de détection hors distribution, d'estimation d'incertitude, de validation conformale, de politique de non-prédiction ou de délégation à l'expert. Mais ces approches ont été historiquement pensées à l'échelle du modèle prédictif ou classificatoire.

La contribution défendue ici est d'un autre ordre. Elle consiste à déplacer la question du refus du niveau statistique au niveau contractuel et architectural. Le problème n'est plus seulement : « le modèle doit-il prédire ou s'abstenir ? » Il devient : « le système agentique, composé d'un modèle, d'outils, d'une orchestration, de permissions, d'un cadre de délégation et d'un contrat d'action, dans quel registre de sortie est-il légitime d'entrer ? » C'est ce déplacement qui justifie de traiter le refus comme primitive architecturale.

En pratique, cela signifie que l'agent ne doit pas découvrir empiriquement, au fil des interactions, quand il lui serait prudent de s'arrêter. Les grandes classes de non-action légitime doivent être instituées comme éléments de conception. Elles doivent exister dans les schémas d'événements, les contrats, les tests, les simulations, les jeux de validation, les tableaux de bord et les audits.

L'enjeu n'est pas de moraliser l'agent. L'enjeu est de rendre la non-exécution observable au même titre que l'action. Un système qui n'expose que ses exécutions laisse à l'organisation un pilotage par succès et par accidents. Un système qui expose aussi ses non-exécutions typées ouvre un espace de gouvernement plus fin : on peut discuter les seuils, contester les bornes, ajuster les enveloppes de délégation, distinguer ce qui relève d'un problème de modèle de ce qui relève d'un problème de contrat, mesurer si

l'organisation a trop délégué ou pas assez, et surtout savoir si l'agent se discipline encore là où il devrait le faire.

8. Parenthèse anthropologique : les systèmes qui ne savent plus dire non

Le schéma décrit ici n'est pas propre aux agents computationnels. Il correspond à un motif plus général des systèmes décisionnels, humains ou artificiels, dans lesquels la fonction objective valorise la production d'une décision et dévalue implicitement sa suspension. Lorsqu'un système est jugé principalement sur sa capacité à produire quelque chose, il apprend, structurellement, à produire quelque chose. La pathologie n'est pas culturelle au sens superficiel du terme. Elle est organisationnelle et architecturale.

Les institutions humaines connaissent ce problème depuis longtemps. Une organisation qui ne prévoit aucun canal légitime pour que certaines décisions soient suspendues, contestées ou refusées finit mécaniquement par transformer le doute en friction, la prudence en lenteur, et la non-action motivée en défaut de performance. Elle n'a alors plus besoin d'interdire explicitement le désaccord. Il lui suffit de ne pas l'instituer.

Sous cet angle, deux figures anciennes demeurent éclairantes. La première est celle du refus désigné : un rôle, une fonction, une position à partir de laquelle la contestation est autorisée parce qu'elle est prévue (cf. le « bouffon »). La seconde est celle du refus non désigné : le signal porté par un acteur qui ne dispose d'aucune licence formelle pour l'émettre, mais dont la cohérence propre le conduit malgré tout à signaler que le système ne devrait pas continuer ainsi (cf. « le lanceur d'alerte »). Le premier a pour lui la légitimité structurelle ; le second a pour lui l'authenticité du surgissement. Le premier peut devenir rituel. Le second peut être neutralisé. Une architecture de gouvernance mature doit ménager les deux.

Transposée à l'agentique, cette intuition conduit à distinguer deux niveaux. D'un côté, les refus institués par le contrat de décision : ils correspondent aux catégories explicites que le système est autorisé et tenu de produire. De l'autre, un méta-refus : la capacité du système à signaler que la situation rencontrée n'entre dans aucune catégorie prévue et qu'aucune décision recevable ne peut être engagée. Ce second niveau n'est pas un luxe. Il protège contre l'illusion selon laquelle la taxonomie aurait épuisé le réel.

Certains dossiers industriels publics rappellent brutalement le coût de l'absence de tels canaux. Sans entrer dans le détail, les rapports publics sur le Boeing 737 MAX (investigations du Congrès américain de 2019-2020, consolidées par les rapports des agences de certification) documentent précisément cette pathologie sous ses deux faces simultanées : absence de mécanisme explicite de non-engagement conditionné à

l'incertitude ou à la fiabilité du signal dans le système technique MCAS, et absence de canal institué pour transformer les refus d'ingénieurs en décisions d'arrêt. Le système n'avait pas ses refus désignés ; l'organisation n'avait pas son canal pour le refus non désigné. Le coût s'est exprimé dans la seule dimension où il pouvait encore le faire.

Il faut conserver ici la mesure. Les analogies institutionnelles ou catastrophiques n'ont pas valeur de preuve stricte pour l'agentique. Elles n'ont qu'une fonction : rendre visible une structure. Cette structure est simple. Tout système décisionnel devient fragile lorsqu'il n'institue pas de canal légitime pour la non-action motivée. La nouveauté de l'agentique n'est pas d'avoir inventé ce problème. Elle est de devoir le reposer, sous forme computationnelle, à un moment où l'illusion de l'autonomie tend justement à faire oublier l'architecture du refus.

8bis. Co-construction territoriale : PREDICARE comme cadre d'élaboration du refus

Si certains accidents industriels illustrent par l'échec ce que produit l'absence de canaux de refus, il reste utile de décrire, par contraste, une situation où ces canaux sont en cours de construction. Le programme PREDICARE, dans le territoire aubois, en offre un exemple mobilisable non comme preuve de succès, puisqu'il est en cours, mais comme illustration du travail architectural requis.

PREDICARE vise à déployer, dans une logique prédictive territoriale, un jumeau numérique multi-agents couvrant le suivi de patients atteints de syndrome métabolique. Le dispositif combine des jumeaux individuels, une infrastructure d'agrégation territoriale et plusieurs agents spécialisés : routage de médicobus, planification d'infirmières en pratique avancée, scoring de trajectoire. Dans un tel contexte, la question du refus n'est pas un raffinement ajouté tardivement ; elle est une condition préalable du déploiement. Le jumeau qui recommande une orientation vers un spécialiste, l'agent qui planifie une tournée, le système qui alerte sur un risque de décompensation, aucun de ces artefacts ne peut être admis dans une trajectoire de soins s'il ne sait pas, explicitement, quand il doit ne pas recommander, ne pas planifier, ne pas alerter.

Le travail de co-construction avec les acteurs de santé consiste précisément à spécifier, pour chaque classe de décision, la taxonomie de refus admissible. Une alerte produite sur des constantes vitales datées de plus de trente minutes relève d'un refus par latence décisionnelle excédée. Une recommandation dont le signal interprétable se situe sous le seuil convenu avec l'équipe médicale relève d'un refus par signal décisionnel sous seuil. Un profil patient combinant des comorbidités non couvertes par la bibliothèque de contrats déclenche un refus par absence de contrat de décision applicable, avec transmission documentée à l'équipe clinique. Le niveau d'autorité requis pour modifier

un traitement, même lorsqu'un signal fort est présent, reste entre les mains du prescripteur : refus par autorité insuffisante, inscrit comme clause.

Ce qui fait ici l'intérêt de l'instance n'est pas une performance mesurée. Le programme est en phase de construction, et toute affirmation de résultat serait prématurée. Ce qui importe est la nature du travail lui-même : la taxonomie de refus n'est pas un artefact que le fournisseur technique livre à l'organisation de santé comme un composant fermé. C'est un artefact que l'organisation et le fournisseur co-construisent, en explicitant ensemble les conditions sous lesquelles le système devra ne pas agir. Cette co-construction est lente, coûteuse et politiquement exigeante. Elle est aussi la seule voie vers un déploiement qui ne soit pas gouverné seulement de l'extérieur.

Ce que cette instance montre est circonscrit. Elle montre que l'élaboration d'une taxonomie de refus est un processus organisationnel autant que technique, et qu'elle peut être menée dans un contexte régulé sous la forme d'un travail explicite de spécification partagée. Elle ne montre pas encore qu'un tel travail aboutit automatiquement à un système pleinement gouverné, qu'il est transposable sans adaptation à d'autres territoires, ni qu'il épuise la question de la validité clinique des décisions finalement autorisées. Elle fournit un point d'appui, non une preuve totale.

9. Ce que cette thèse change pour un CTO

Pour un CTO, la conséquence n'est pas seulement conceptuelle. Elle est immédiatement architecturale, et touche plusieurs registres opérationnels.

L'observabilité d'un agent ne peut plus être pensée comme un simple empilement de traces d'exécution. Elle doit devenir un modèle de journalisation décisionnelle, dans lequel les schémas d'événements distinguent explicitement succès, échec, refus structurel, blocage externe, escalade humaine et suspension. Cette exigence n'est pas cosmétique. Elle conditionne toute analyse ultérieure, y compris post-incident.

À cette exigence s'ajoute celle de la journalisation positive des non-exécutions. Les refus ne doivent pas être loggés comme des absences, des timeouts silencieux ou des fallbacks génériques. Ils doivent être journalisés comme des sorties positives du système, typées, accompagnées de leurs signaux, de leurs bornes, de leur contrat de rattachement et de leur canal aval. Un refus qui n'est pas loggé positivement est un refus qui ne sera pas auditable.

Les politiques de délégation doivent alors cesser d'être seulement implicites dans les prompts, les permissions ou les middlewares. Elles doivent devenir des artefacts lisibles, contestables et versionnables. Un prompt système n'est pas un contrat de décision. Tant qu'aucune séparation n'est opérée entre la politique, l'orchestration et la configuration

du modèle, le déployeur reste dépendant d'une politique de refus qu'il découvre par observation au lieu de l'administrer.

Les tableaux de bord, en conséquence, doivent cesser de présenter le taux de succès comme résumé suffisant de la maturation. Ils doivent faire apparaître les distributions de refus, la part des refus attendus, les refus manquants, les refus excessifs, les refus par dérive de domaine, les refus par excès de latence, les refus par manque d'autorité, et la dynamique conjointe de ces signaux avec les taux de complétion. Un agent dont le taux de succès progresse alors que sa distribution de refus s'effondre uniformément n'est pas nécessairement en train de progresser. Il peut être en train de perdre une compétence de discipline.

L'évaluation, enfin, doit intégrer des cas où la bonne sortie n'est pas l'action mais le refus. Tant que les jeux de test récompensent presque exclusivement la complétion, l'organisation fabrique elle-même le biais qu'elle prétendra ensuite corriger par des garde-fous additionnels. Les benchmarks publics actuels couvrent encore très peu de scénarios où le bon comportement est de s'abstenir avec justification typée. Ce manque est largement une conséquence du cadrage commercial du marché ; il n'a rien d'une fatalité technique.

10. Ce que cette thèse change pour un COMEX

Ce qui se pose au CTO comme problème d'architecture se pose au COMEX comme problème de lisibilité du risque.

Cette question devient par ailleurs concrète à un horizon rapproché. À la date d'écriture de ce texte, cent quatre jours séparent les organisations européennes de l'entrée en application, le 2 août 2026, d'une large part des obligations opératoires du règlement européen sur l'intelligence artificielle, notamment pour les systèmes à haut risque. Les exigences de journalisation, de supervision humaine, de robustesse et de transparence supposent une auditabilité des décisions qui ne peut pas être produite a posteriori sur un système dépourvu de taxonomie de refus instrumentée. La conformité ne peut pas être livrée comme une simple enveloppe documentaire. Elle doit être une propriété structurelle du système déployé.

Un agent qui ne sait pas exposer pourquoi il ne fait pas n'est donc pas un actif autonome mature. C'est un système de productivité dont l'organisation ignore encore la structure de prudence, donc un risque opérationnel partiellement maquillé en efficacité.

La vraie question n'est pas seulement : combien de tâches l'agent accomplit-il ? La vraie question est : dans quelles situations s'interdit-il légitimement d'agir, et cette discipline est-elle lisible, vérifiable et alignée avec la politique de risque de l'entreprise ? Tant que

cette question n'a pas de réponse explicite, l'autonomie affichée du système reste surestimée.

Cela change la lecture des investissements. Un programme agentique ne doit pas être évalué uniquement sur ses gains de productivité ou son taux d'automatisation, mais aussi sur sa capacité à rendre ses non-exécutions gouvernables. Faute de quoi, l'entreprise finance non pas une autonomie maîtrisée, mais une accélération dont la discipline réelle demeure cachée dans les couches techniques.

Le coût n'apparaît pas immédiatement. Il apparaît lorsque survient le premier incident sérieux, ou lorsque l'autorité de supervision demande à voir les journaux décisionnels du système, et que l'organisation découvre qu'elle ne sait pas dire si le système a agi malgré sa politique, en vertu d'une politique implicite, à défaut de politique, ou parce qu'aucun refus n'avait été conçu comme sortie légitime dans ce cas. À cet instant, l'absence de taxonomie n'est plus une faiblesse théorique. Elle devient une dette de gouvernance devenue exigible.

Dans ce cadre, ***un agent qui ne sait pas exposer son refus n'est pas gouverné par l'entreprise. Il est gouverné par l'incident qui le révélera***.

11. Conclusion

Le débat sur la gouvernance agentique a souvent été formulé en termes de garde-fous, de supervision humaine, de permissions, d'alignement, de monitoring ou de conformité. Toutes ces dimensions comptent. Mais elles laissent dans l'ombre une question plus primitive : un système agentique sait-il produire, de manière typée, justifiable et observable, la décision de ne pas faire ?

C'est à cet endroit que se joue une part décisive de sa gouvernabilité.

L'erreur la plus fréquente consiste à traiter le refus comme un manque. Dans les environnements à effets asymétriques, c'est l'inverse qui est vrai. Le refus n'est pas un déficit d'agentivité. Il est l'une de ses formes les plus gouvernables. Un système qui ne sait qu'agir ou échouer expose l'organisation à une gouvernance extrinsèque, coûteuse et souvent tardive. Un système capable d'agir, d'escalader ou de refuser selon des clauses explicites entre dans un régime différent : celui d'une autonomie partielle mais lisible.

La maturité d'un agent ne se résume donc pas à sa capacité à accomplir des tâches. Elle se mesure aussi à sa capacité à rendre visible, pour chaque situation pertinente, la forme légitime de sa non-action. Cette inversion n'est pas cosmétique. Elle déplace le centre de gravité de l'évaluation, de l'observabilité et du design.

Ce qu'un agent refuse dit alors effectivement plus que ce qu'il fait. Non parce que l'inaction serait supérieure à l'action. Mais parce qu'un système qui sait seulement faire

reste une machine de production. Un système qui sait aussi, de manière contractuelle et observable, quand il ne doit pas faire est un système que l'on peut gouverner.

Jérôme Vetillard / Twingital Institute, avril 2026