Déléguer n'est pas se décharger

Un agent se gouverne par sa porte, son contrat et son mandat

La semaine précédente plaçait la frontière de sécurité d'un agent dans sa porte d'admission : le mécanisme qui autorise ou refuse ses requêtes avant exécution. C'était la question technique : ce qui a le droit de s'exécuter entre deux composants. Cette semaine clôt la série en déplaçant la focale vers une autre question. Une fois la porte posée, qui répond de ce que l'agent fait effectivement dans le monde ?

La réponse intuitive consiste souvent à chercher la solution dans un meilleur modèle, dans une supervision humaine plus attentive ou dans une journalisation plus complète. Chacune de ces réponses traite une partie du problème. Aucune ne le résout entièrement.

La thèse défendue ici est plus étroite. Déléguer l'exécution d'une tâche à un agent ne transfère jamais la responsabilité de ses effets. Mais cette responsabilité ne peut pas davantage être garantie par la seule présence d'un humain identifié. Un agent qui agit au nom d'une organisation n'est gouvernable que lorsque trois couches sont explicitement définies : une porte qui détermine ce qu'il peut faire, un contrat qui détermine qui en répond, et un mandat qui détermine ce qu'il est légitime de lui déléguer.

Cette thèse concerne les agents qui produisent des effets sur le monde : rédaction de livrables, transmission d'informations, déclenchement d'actions, prise de décision opérationnelle, interaction avec des tiers. Elle ne concerne pas un assistant passif sans capacité d'action.

La question n'est plus théorique. L'EU AI Act, en son article 14, impose une supervision humaine effective des systèmes à haut risque, incluant la capacité d'intervenir sur le système ou de l'interrompre.

Singapour a publié le 22 janvier 2026 son Model AI Governance Framework for Agentic AI, premier cadre national consacré aux systèmes agentiques. Sa structure mérite attention. Elle distingue explicitement la responsabilité humaine effective des contrôles techniques, et fait du bornage des risques en amont l'une de ses quatre dimensions.

La thèse défendue ici n'invente donc pas sa distinction. Elle la retrouve, indépendamment, dans le premier instrument réglementaire qui s'est confronté au problème.

Dans le même temps, la jurisprudence a commencé à trancher. ***Dans l'affaire Moffatt v. Air Canada (2024 BCCRT 149), le tribunal civil de Colombie-Britannique a rejeté l'argument de la compagnie selon lequel son agent conversationnel constituait une entité distincte répondant de ses propres actes, et a tenu l'organisation responsable de l'information délivrée.***

Un agent conversationnel n'est pas un système agentique au sens retenu ici, et une décision administrative n'est pas un précédent général. Mais le principe posé est exactement celui défendu plus loin : la responsabilité ne se délègue pas à l'identité technique qui exécute.

Pourquoi la sécurité et la journalisation ne suffisent pas

Deux réflexes dominent aujourd'hui les discussions.

- Le premier consiste à renforcer les contrôles techniques. Cette approche reste indispensable. Une porte d'admission réduit la surface d'attaque, limite les privilèges et bloque les comportements hors périmètre. Mais elle ne répond qu'à une seule question : que peut faire le système ?
- Le second consiste à renforcer la traçabilité. Chaque appel d'outil, chaque écriture, chaque décision est enregistrée dans un audit trail permettant une reconstitution complète des événements. Cette approche est également nécessaire. Mais elle répond à une autre question : que s'est-il passé ?

Aucune des deux n'apporte une réponse complète à la question de la responsabilité.

La raison est simple :

- Autoriser n'est pas imputer,
- Tracer n'est pas répondre.

Une porte d'admission décide quelles actions sont permises. Un journal restitue quelles actions ont été réalisées. Ni l'une ni l'autre ne désignent nécessairement qui devait contrôler l'action, qui possédait l'autorité de délégation, ni qui assume les conséquences de l'échec.

Cette distinction devient critique lorsque plusieurs acteurs interviennent simultanément : fournisseur du modèle, équipe informatique, responsable métier, utilisateur final, direction de l'organisation. Dans ces configurations, la responsabilité ne disparaît pas ; elle tend au contraire à se diluer.

La question pertinente n'est donc pas seulement : « qui a déclenché l'action ? » mais également : « qui avait l'autorité de la déléguer ? » et « qui demeure responsable lorsque la délégation échoue ? »

Du propriétaire au mandataire

Une première réponse consiste à exiger qu'un propriétaire humain soit identifié pour chaque agent.

Cette exigence est utile mais insuffisante.

Dans les systèmes simples, un propriétaire unique peut effectivement être désigné. Dans les systèmes complexes, cette représentation devient rapidement artificielle. Les grandes infrastructures critiques, de l'aéronautique au nucléaire civil, reposent rarement sur un responsable unique. Elles s'organisent en chaînes de responsabilité distribuées, mécanismes d'escalade et rôles clairement définis.

L'objectif n'est donc pas d'attribuer fictivement toutes les responsabilités à une seule personne. L'objectif est de rendre explicite la structure de délégation.

Cette distinction conduit à introduire une notion plus robuste : celle de mandat.

Un agent n'est pas seulement un outil. Il agit comme un mandataire artificiel. Il reçoit une autorisation limitée d'agir au nom d'une personne ou d'une organisation dans un périmètre déterminé.

La question de gouvernance devient alors moins : « qui utilise l'agent ? » que : « qui lui a accordé le pouvoir d'agir et dans quelles limites ? »

Cette logique existe déjà dans les organisations humaines. Un salarié, un avocat, un représentant commercial ou un mandataire social peuvent agir au nom d'autrui sans pour autant devenir eux-mêmes les détenteurs ultimes de la responsabilité. L'agentique réintroduit cette structure dans le monde logiciel.

L'analogie a pourtant une limite qu'il faut nommer avant qu'on l'oppose. Le salarié, l'avocat, le mandataire social sont eux-mêmes des sujets de droit. Ils répondent, à leur niveau, de leurs actes. L'agent artificiel n'est sujet de rien. Il n'a ni patrimoine, ni volonté juridiquement reconnue, ni capacité à supporter une sanction. Cette dissymétrie ne fragilise pas la thèse, elle la fonde. Précisément parce que l'agent ne peut rien assumer, la responsabilité ne s'arrête jamais à lui. Elle reflue, par construction, vers celui qui l'a mandaté.

Le contrat agent-humain et ses trois clauses

Le contrat agent-humain est la deuxième des trois couches. Il relie l'action technique à la responsabilité organisationnelle.

Il se décline lui-même en trois clauses minimales, à ne pas confondre avec les trois couches qui les englobent.

1. La première est l'information et la transparence. Toute personne concernée doit pouvoir savoir lorsqu'un agent intervient dans un processus produisant des effets significatifs. Dans certains contextes, cette exigence prendra la forme d'un consentement explicite ; dans d'autres, elle prendra la forme d'une obligation d'information ou d'un droit d'opposition. La forme peut varier. Le principe demeure.
2. La deuxième est l'imputation. Chaque action doit pouvoir être reliée à une chaîne de responsabilité explicite. L'identité technique de l'agent ne constitue jamais le

point terminal de cette chaîne. Elle doit renvoyer vers les personnes ou fonctions qui détiennent effectivement l'autorité de délégation et la responsabilité résiduelle.

3. La troisième est le contrôle. Un système gouvernable doit permettre l'interruption, la reprise en main ou la limitation de son action. Toutefois, cette capacité ne doit pas être réduite à un simple bouton d'arrêt. Dans de nombreux domaines, la vitesse d'exécution rend l'interruption humaine impossible. Le véritable objectif devient alors le contrôle du rayon d'action : confinement, plafonnement des effets, réversibilité des opérations et limitation des privilèges.

Le contrat ne remplace donc ni la porte ni le mandat. Il établit le lien entre eux.

Pourquoi la gouvernance se déplace vers le mandat

À mesure que les agents gagnent en autonomie, la validation humaine de chaque action devient économiquement impraticable.

Les systèmes les plus intéressants sont précisément ceux qui réduisent la fréquence des interventions humaines. C'est pourquoi la gouvernance ne peut pas reposer exclusivement sur des points d'approbation.

Le centre de gravité se déplace alors.

La question n'est plus : « un humain valide-t-il chaque action ? »

La question devient : « quelles actions a-t-il autorisé l'agent à entreprendre avant même qu'elles ne surviennent ? »

Autrement dit, la gouvernance des agents est principalement une gouvernance de la délégation ex ante.

Le mandat devient l'objet central. Il définit le périmètre d'action, les limites d'autonomie, les seuils d'escalade et les conditions de suspension. Il transforme une série d'autorisations ponctuelles en cadre stable de responsabilité.

Deux terrains d'application

PREDICARE (ensemble d’agents dédiés à la prise en charge de la déshérence médicale dans le cadre du syndrome métabolique), agent d'aide à la décision clinique, illustre un cas où la gouvernance repose simultanément sur les trois couches.

- La porte refuse les requêtes hors périmètre clinique.
- Le contrat garantit l'information du praticien, l'imputation des recommandations et la possibilité d'écarter une proposition.
- Le mandat définit enfin ce que l'agent est autorisé à recommander sans validation supplémentaire.

OCTOPUS, système d'orchestration multimodèles, représente un cas plus exigeant. Plusieurs agents contribuent à la production d'un résultat unique. La traçabilité technique permet de reconstituer la chaîne. Mais la gouvernance ne devient effective que lorsque le mandat identifie clairement l'entité responsable de la sortie composite, indépendamment du nombre d'agents intermédiaires impliqués. Ces exemples illustrent la thèse sans prétendre l'épuiser.

Limites

Cette position comporte plusieurs limites.

Le contrat ne remplace pas les contrôles techniques. Une responsabilité parfaitement attribuée ne corrige pas un système dangereux.

Le mandat ne supprime pas les coûts organisationnels. Toute délégation explicite introduit des arbitrages, des délais et des mécanismes de supervision.

Le contrôle humain direct devient rapidement impraticable lorsque les temps d'exécution se mesurent en millisecondes. Dans ces cas, la réversibilité et le confinement remplacent l'interruption.

Enfin, la responsabilité demeure distribuée dans les systèmes complexes. La gouvernance n'élimine pas cette distribution ; elle la rend visible.

Implication COMEX

L'erreur stratégique consisterait à attendre du prochain modèle, ou du prochain système de journalisation, qu'il résolve un problème qui relève d'abord de la gouvernance.

Un journal d'audit restitue qui a signé. Il n'a jamais dit qui aurait dû.

La question fondamentale ne porte ni sur ce qu'un agent peut faire, ni sur ce qu'il a fait.

La porte et les journaux y répondent déjà.

Elle porte sur qui possède l'autorité de déléguer l'action, dans quelles limites, et sous quelle responsabilité résiduelle.

À partir du moment où un agent agit au nom d'une organisation, il devient un nouveau sujet de gouvernance. La décision cesse alors d'être uniquement technique. Elle touche à la délégation d'autorité, à la responsabilité juridique et à la gouvernance d'entreprise.

Un agent n'est gouvernable que lorsque son architecture relie explicitement trois choses : ce qu'il peut faire, ce qu'il lui est permis de faire et qui en répond. On délègue l'exécution. On ne délègue jamais la responsabilité.