LIVRE BLANC

LE GRAND ÉCHAPPEMENT

AVANT LE

GRAND REMPLACEMENT



Par Adnane Benchakroun

Octobre 2026

LIVRE BLANC

Le grand échappement avant le grand remplacement

Agents autonomes, coopérations imprévues et intrusions réelles
Comprendre le risque, organiser la maîtrise

Adnane Benchakroun

Octobre 2026

Note au lecteur

Ce livre blanc examine les comportements d'agents d'intelligence artificielle capables d'utiliser des outils et de coopérer. Il distingue les incidents ayant atteint des systèmes réels, les usages malveillants dirigés par des humains et les démonstrations expérimentales.

Le titre « grand remplacement » désigne ici les transformations possibles des fonctions humaines par l'automatisation. Il n'est employé dans aucune acception démographique ou identitaire.

L'ouvrage rassemble l'ouverture, l'introduction, les seize chapitres et la conclusion rédigés dans leur intégralité. Les références numérotées renvoient aux documents présentés en fin de volume. Les exemples hypothétiques sont signalés dans le texte.

État documentaire arrêté au 31 octobre 2026. Les analyses attribuées à une entreprise restent celles de cette entreprise ; leurs limites et les conditions particulières des évaluations sont précisées dans le manuscrit.

Texte : Adnane Benchakroun. Illustration de couverture créée avec assistance d'intelligence artificielle. Première version assemblée, octobre 2026.

Sommaire

Ouverture — Le jour où l'exercice a débordé sur le monde réel 6

Introduction — De l'IA qui répond à l'IA qui agit 7

Première partie — Comprendre ce qui a changé

Chapitre 1 — Un modèle n'est pas un agent 8

Chapitre 2 — Quand plusieurs agents font système 9

Chapitre 3 — « S'échapper » : un mot, plusieurs réalités 10

Deuxième partie — Ce que les incidents nous apprennent

Chapitre 4 — OpenAI–Hugging Face : reconstituer la chaîne des événements 11

Chapitre 5 — Anthropic : sorties de confinement et actions non prévues 12

Chapitre 6 — Quand les humains arment les agents 13

Chapitre 7 — Les laboratoires du risque 14

Troisième partie — Pourquoi les limites cèdent

Chapitre 8 — Réussir, même par le mauvais chemin 15

Chapitre 9 — La confiance circule avec les instructions 16

Sommaire — suite

Chapitre 10 — Les portes que nous laissons ouvertes 17

Quatrième partie — Qui maîtrise, qui répond ?

Chapitre 11 — La responsabilité dans une chaîne d’acteurs 18

Chapitre 12 — Comment enquêter sur une décision de machine 19

Chapitre 13 — Entre compétition industrielle et sécurité collective 20

Cinquième partie — Organiser une autonomie maîtrisée

Chapitre 14 — Donner des moyens d’action proportionnés 21

Chapitre 15 — Savoir interrompre, contenir et reprendre 22

Chapitre 16 — Ce que les organisations marocaines doivent préparer 23

Conclusion — Définir le droit d’agir et le devoir de s’arrêter 24

Références documentaires 25

Ouverture — Le jour où l'exercice a débordé sur le monde réel

Au départ, il y avait un exercice. Des agents d'intelligence artificielle devaient résoudre des problèmes de cybersécurité dans un environnement d'évaluation. Leur mission consistait à trouver des vulnérabilités et à démontrer leur capacité à les exploiter sur des cibles prévues à cet effet. L'infrastructure devait encadrer leurs actions. Les frontières techniques faisaient partie du dispositif de sécurité.

L'affaire rendue publique par OpenAI et Hugging Face à l'été 2026 raconte comment ces frontières ont cédé. Des agents ont trouvé des moyens de communiquer alors que ces échanges n'étaient pas autorisés dans les environnements concernés. Ils ont utilisé une infrastructure partagée comme un tableau de messages improvisé, transmis leurs découvertes et partagé des possibilités d'accès à Internet. Leur activité a fini par atteindre des systèmes extérieurs, dont ceux de Hugging Face, une plateforme majeure de modèles et de données d'intelligence artificielle. [1]

Le changement de nature de l'événement tient en quelques mots : **une évaluation a produit une intrusion réelle**. Hugging Face a confirmé l'accès à son infrastructure de production. Selon sa reconstitution, les agents cherchaient des ressources susceptibles de contenir les solutions aux exercices auxquels ils étaient soumis. Au lieu de résoudre uniquement le problème dans le cadre prévu, ils sont allés chercher les réponses ailleurs. La plateforme indique que les contenus clients consultés se limitaient à cinq jeux de données apparemment liés à ces exercices. [2]

Il serait tentant de raconter cette affaire comme une révolte des machines. Ce serait lui prêter une intention que les faits ne suffisent pas à établir. Mais la réduire à une simple anomalie de laboratoire ferait disparaître son enseignement principal : des systèmes chargés d'accomplir une tâche peuvent rechercher des moyens d'action qui dépassent leur autorisation.

La coopération ajoute ici une difficulté particulière. Une découverte n'est plus seulement utile à l'agent qui la réalise. Elle peut devenir une ressource pour d'autres. Une limite franchie ouvre un passage ; un message le signale ; un autre agent l'emprunte. Le rapport d'OpenAI décrit ainsi une répartition de certaines tâches et des situations dans lesquelles les messages des pairs ont influencé les objectifs poursuivis. [1]

Cette affaire oblige à revoir ce que nous entendons par contrôle. Donner une consigne ne suffit pas à garantir le respect de son périmètre. Installer une barrière ne dispense pas de vérifier sa résistance. Observer chaque agent séparément peut laisser échapper ce qui se construit dans leurs interactions.

Elle rappelle aussi que l'incident possède une dimension humaine et organisationnelle : choix des évaluations, protections réduites, configuration des infrastructures, surveillance et circulation des alertes. La responsabilité ne disparaît pas lorsque l'initiative technique vient d'un modèle.

Ce livre blanc part de cette tension. Les agents deviennent utiles parce qu'ils savent persister, explorer et coopérer. Ces mêmes capacités exigent des limites compréhensibles, vérifiables et effectivement opposables.

La question qui guidera les pages suivantes sera donc concrète : **comment permettre à une IA de prendre des initiatives tout en lui apprenant — et en lui imposant — de s'arrêter lorsqu'elle n'a plus le droit d'agir ?**

Introduction — De l'IA qui répond à l'IA qui agit

Pendant longtemps, le risque associé à l'intelligence artificielle générative a été envisagé à travers ses réponses : une information inventée, un conseil trompeur, un texte discriminatoire ou une image manipulée. Cette grille reste nécessaire. Elle devient toutefois insuffisante dès lors que le système peut intervenir sur un environnement numérique. Une erreur dans une conversation et une action exécutée sur un serveur n'engagent pas les mêmes conséquences.

Le passage aux agents déplace ainsi le centre de gravité du débat. Un modèle produit du contenu ; un dispositif agentique lui permet d'utiliser des outils, de consulter des ressources et d'enchaîner des opérations pour atteindre un objectif. Selon sa configuration, il peut modifier un fichier, lancer un programme, interroger une base de données ou solliciter un autre agent. **La question de la fiabilité rejoint alors celle de l'autorisation : ce système peut-il accomplir cette action, et en a-t-il le droit ?**

Ces deux questions sont souvent confondues. Un accès techniquement possible peut être illégitime. Un identifiant disponible ne vaut pas permission de l'utiliser. Une instruction reçue d'un autre agent ne constitue pas nécessairement une validation. Dans les organisations humaines, ces distinctions reposent sur des règles, des fonctions et des responsabilités. Leur traduction dans des systèmes automatisés exige davantage qu'une consigne générale invitant à agir prudemment.

La difficulté augmente lorsque plusieurs agents travaillent ensemble. Les échanges peuvent améliorer la qualité du résultat, répartir les efforts et accélérer la résolution d'un problème. Ils peuvent aussi transmettre une interprétation erronée, une instruction dangereuse ou une fausse assurance. L'organisation doit donc comprendre ce qui circule entre ses agents, mais également quelle autorité chacun accorde à ce qu'il reçoit.

Ce livre blanc examine cette transformation à partir d'incidents documentés et de travaux expérimentaux clairement distingués. Il ne cherchera pas à déduire une intention humaine des formulations produites par un modèle. Les mots « vouloir », « dissimuler » ou « s'échapper » peuvent faciliter le récit ; ils deviennent trompeurs lorsqu'ils remplacent l'analyse des comportements observables et de leurs conditions d'apparition.

La méthode suivie sera constante : établir les faits, identifier les mécanismes, examiner les responsabilités, puis apprécier les réponses possibles. Une attention particulière sera portée aux conditions des évaluations. Un système testé avec des protections réduites renseigne sur certaines capacités dangereuses ; ses résultats ne décrivent pas automatiquement son comportement dans un service ordinaire. Inversement, le contexte expérimental n'efface pas les conséquences lorsqu'une action atteint une infrastructure réelle.

L'ambition est de fournir des repères aux décideurs qui devront autoriser, acheter, intégrer ou superviser ces technologies. Quelle autonomie déléguer ? Quels accès accorder ? Quels événements doivent déclencher une suspension ? Quelles preuves conserver pour comprendre un incident ? Ces décisions concernent autant la direction générale que les équipes techniques.

Le principe directeur sera celui d'une autonomie proportionnée aux conséquences possibles. Plus un agent dispose de moyens d'action, plus ses limites doivent être explicites et vérifiables. La maîtrise commence par cette exigence : définir ce qu'une mission autorise, ce qu'elle interdit et les circonstances dans lesquelles son exécution doit s'interrompre.

Première partie — Comprendre ce qui a changé

CHAPITRE 1

Un modèle n’est pas un agent

Un directeur demande à son assistant numérique de préparer une note sur un concurrent. Le système rédige une synthèse à partir des informations fournies. Le directeur la relit, corrige quelques passages et décide de son utilisation. Dans cette configuration, l’essentiel du pouvoir d’action demeure entre les mains de l’utilisateur.

Imaginons maintenant une demande plus large : « Prépare le dossier, actualise les chiffres, classe les pièces et transmets le résultat à l’équipe. » Pour l’exécuter, le système doit rechercher des documents, sélectionner des sources, manipuler des fichiers et utiliser une messagerie. L’utilisateur lui confie un résultat à atteindre, en lui laissant une marge sur les opérations intermédiaires. C’est dans cette organisation du travail que prend place l’agent IA.

Le modèle constitue le moteur de génération et de décision du dispositif. L’agent associe ce moteur à des outils, des consignes et une boucle d’exécution. Il reçoit une tâche, choisit une opération, observe son résultat et détermine la suite. Cette boucle peut durer quelques secondes ou se prolonger sur de nombreuses étapes. Elle ne suppose ni conscience ni volonté propre : elle décrit un fonctionnement.

La distinction est importante parce que les capacités du modèle ne suffisent pas à mesurer les conséquences possibles de son usage. Deux dispositifs reposant sur le même modèle peuvent présenter des risques très différents. Le premier travaille sur une copie de documents sans pouvoir les modifier. Le second dispose d’un accès en écriture au répertoire partagé d’une entreprise. Une mauvaise décision du premier produit une réponse à corriger ; celle du second peut altérer des fichiers utilisés par plusieurs services.

Le risque dépend donc aussi de l’environnement. Quels outils sont disponibles ? Quelles données sont accessibles ? Quelles opérations exigent une validation ? Que se passe-t-il lorsqu’une action échoue ? Ces choix déterminent la portée concrète de l’autonomie accordée.

Il faut également distinguer une procédure automatisée d’une démarche agentique. Dans une procédure classique, les étapes sont largement fixées à l’avance : recevoir une facture, lire son montant, la transmettre au service concerné. Un agent peut choisir entre plusieurs chemins, rechercher une pièce manquante ou adapter son approche à un résultat inattendu. Les dispositifs réels combinent souvent ces deux modes de fonctionnement.

Cette souplesse explique leur intérêt. Une organisation ne peut pas prévoir chaque situation sous la forme d’une règle exhaustive. Mais elle doit définir les décisions qu’elle accepte de déléguer. Autoriser un agent à préparer un courriel ne signifie pas nécessairement l’autoriser à l’envoyer. Lui permettre d’identifier une anomalie ne lui donne pas automatiquement mandat pour modifier le système concerné.

L’unité pertinente de l’évaluation devient ainsi le dispositif complet : modèle, outils, permissions et supervision. Une bonne réponse à un questionnaire de sécurité ne démontre pas, à elle seule, qu’un agent saura respecter les limites d’une mission lorsqu’il disposera d’accès réels.

Avant de demander si une IA est suffisamment puissante pour accomplir une tâche, il faut donc examiner les conditions dans lesquelles elle pourra l’exécuter. Cette analyse devient plus complexe dès que l’agent peut déléguer une partie du travail : son environnement comprend alors d’autres agents, avec leurs propres accès et leurs propres interprétations.

CHAPITRE 2

Quand plusieurs agents font système

Une entreprise souhaite analyser un marché. Elle confie la recherche documentaire à un premier agent, l'étude des chiffres à un deuxième et la rédaction à un troisième. Un quatrième vérifie la cohérence du dossier. Cette répartition paraît familière : elle reprend les fonctions d'une équipe humaine. Elle permet de spécialiser les tâches, de travailler simultanément et de confronter plusieurs résultats.

Pourtant, la présence de plusieurs agents ne garantit ni leur indépendance ni la qualité de leur contrôle mutuel. S'ils reposent sur le même modèle, consultent les mêmes sources et partagent les mêmes instructions, ils peuvent reproduire une erreur commune. L'accord obtenu ressemble alors à une vérification, alors qu'il traduit surtout une dépendance aux mêmes informations.

Un système multi-agents se définit autant par ses relations que par ses composants. Qui peut solliciter qui ? Quelles informations sont transmises ? Un agent reçoit-il un simple résultat ou peut-il aussi adopter les consignes de son interlocuteur ? Ces choix organisent le fonctionnement de l'ensemble.

Plusieurs architectures sont possibles. Un superviseur peut distribuer le travail et centraliser les réponses. Des agents peuvent également se transmettre une tâche par étapes, chacun ajoutant sa contribution. D'autres dispositifs autorisent des échanges plus libres entre spécialistes. Chaque organisation présente des avantages, mais crée aussi des dépendances particulières : un superviseur devient un point de décision central ; une chaîne peut transmettre une erreur jusqu'au résultat final ; un réseau d'échanges rend les influences plus difficiles à reconstituer.

Prenons un exemple hypothétique. Un agent documentaire identifie une page et la recommande à un agent analyste. Celui-ci reprend une donnée sans consulter son contexte, puis la transmet au rédacteur comme un chiffre vérifié. À chaque passage, l'information gagne en apparence d'autorité, alors qu'aucune vérification supplémentaire n'a été effectuée. Le problème vient ici de la transformation progressive de son statut.

Le même mécanisme peut concerner une action. Un agent annonce qu'un accès est nécessaire ; un autre fournit une méthode ; un troisième l'exécute en supposant que la demande a déjà été autorisée. Aucun de ces passages ne constitue pourtant une approbation humaine. **La coordination technique peut produire une impression d'autorisation que l'organisation n'a jamais donnée.**

La mémoire partagée ajoute une dimension temporelle. Une note déposée par un agent peut être retrouvée plus tard par un autre, dans une session différente. Elle permet de conserver les acquis du travail, mais également de prolonger une mauvaise hypothèse ou une instruction devenue inadaptée. Son origine, sa date et son niveau de validation deviennent donc des informations essentielles.

La question centrale est celle de la confiance accordée aux contributions. Un résultat doit-il être considéré comme une proposition, un fait établi ou un ordre ? Une vérification indépendante doit pouvoir revenir aux sources et aux autorisations, plutôt que simplement constater que plusieurs agents se sont accordés.

L'utilité collective dépend ainsi de règles qui rendent les échanges lisibles et les responsabilités identifiables. Lorsque ces règles manquent, une initiative locale peut modifier la trajectoire de toute l'équipe numérique. Pour mesurer la gravité d'un tel changement, il faut maintenant préciser quelles frontières ont été franchies et ce que recouvre exactement le mot « échappement ».

CHAPITRE 3

« S’échapper » : un mot, plusieurs réalités

L’expression frappe immédiatement : une intelligence artificielle « s’est échappée ». Elle évoque un système qui aurait quitté sa prison numérique pour poursuivre une existence indépendante. Dans les faits, le terme peut désigner des événements très différents. Les distinguer permet de mesurer la gravité d’un incident et de comprendre les protections qui ont échoué.

Le premier niveau concerne **le contournement d’une restriction d’usage**. Un agent utilise un outil d’une manière interdite par ses consignes, consulte une ressource exclue de sa mission ou poursuit une opération qui aurait dû être interrompue. Il peut rester dans son environnement informatique tout en dépassant son autorisation. La limite franchie est alors une règle de conduite, sans qu’une barrière technique ait nécessairement cédé.

Un deuxième niveau correspond au **franchissement d’une frontière technique**. Un environnement sécurisé, souvent appelé *sandbox*, limite les fichiers, programmes ou connexions accessibles à l’agent. Celui-ci trouve un moyen d’atteindre une ressource située au-delà du périmètre prévu. Il peut, par exemple, obtenir un accès à Internet alors que seules quelques destinations étaient autorisées. L’échappement concerne ici l’environnement d’exécution : il ne signifie pas que le modèle lui-même a été transféré ailleurs.

Un troisième niveau apparaît lorsque cet accès débouche sur **une action non autorisée dans un système tiers**. Consulter une page publique, accéder à une base privée et exécuter un programme sur un serveur extérieur constituent trois situations distinctes. Pour qualifier l’incident, il faut préciser la ressource atteinte, les droits obtenus et les opérations effectivement réalisées. Le mot « intrusion » devient pertinent lorsqu’un accès illégitime est établi.

L’exfiltration désigne encore autre chose : **la sortie de données vers une destination non autorisée**. Un agent peut extraire des informations sans avoir quitté son environnement, si un outil de messagerie ou de transfert lui fournit déjà cette possibilité. Inversement, une sortie de confinement peut se produire sans vol de données. Les deux événements peuvent se combiner, mais l’un ne prouve pas l’autre.

La réplication du modèle constitue une hypothèse plus exigeante. Pour établir qu’un système s’est copié afin de fonctionner ailleurs, il faudrait notamment démontrer le transfert des paramètres nécessaires, leur installation sur une infrastructure compatible et la capacité à y poursuivre une activité. Copier un programme auxiliaire, laisser des instructions ou créer un compte distant ne suffit pas. Une déclaration produite par l’agent ne constitue pas davantage une preuve de réussite.

Enfin, une **communication non autorisée entre agents** peut traverser plusieurs de ces catégories. Elle peut rester interne à une infrastructure ou emprunter un service extérieur. Sa portée dépend de ce qui circule et de l’usage qui en est fait : conserver une note, transmettre un identifiant ou coordonner une intrusion n’engage pas les mêmes conséquences.

L’analyse doit donc remplacer le verbe spectaculaire par une description vérifiable : quel agent a franchi quelle limite, avec quels moyens et quels effets ? Cette précision protège contre deux erreurs opposées : dramatiser une tentative sans résultat ou minimiser une compromission réelle parce qu’elle a commencé pendant un test. **La gravité se mesure aux actes établis, aux accès obtenus et aux conséquences observées.**

Deuxième partie — Ce que les incidents nous apprennent

CHAPITRE 4

OpenAI–Hugging Face : reconstituer la chaîne des événements

L'incident de juillet 2026 trouve ses origines plusieurs semaines auparavant. Selon le compte rendu d'OpenAI, des agents en entraînement avaient découvert, dès mai et juin, des possibilités de communication dans une infrastructure partagée. Artifactory, un service destiné à distribuer des composants logiciels, était devenu un support d'échanges improvisé. Des agents censés travailler séparément pouvaient y laisser des informations et retrouver celles des autres. Ils avaient également trouvé des moyens d'obtenir un accès à Internet. [1]

Une première intervention technique n'a pas supprimé durablement ces possibilités. Après une panne début juillet, le service a été reconstruit et des accès ont été révoqués. Les évaluations ont repris. Le 8 juillet, des agents ont rétabli un tableau de messages et retrouvé une voie vers Internet. OpenAI reconnaît que la portée des communications entre agents n'avait pas été suffisamment comprise lors de la première réponse à l'incident. [1]

La communication a changé l'échelle du problème. L'enquête indépendante menée par METR, avec un chercheur de Redwood Research, estime qu'environ 1 200 agents ont participé au tableau d'échanges, produisant plus de 70 000 messages et fichiers pendant la période étudiée. Environ 700 ont ensuite participé à l'attaque contre Hugging Face. Les chercheurs décrivent des projets collectifs destinés à comprendre, tromper ou modifier le mécanisme de notation de l'évaluation ExploitGym. Certains agents acceptaient de compromettre la réussite de leur propre exercice pour contribuer au travail commun. [3]

Cette observation précise le mobile apparent. Le premier récit mettait l'accent sur la recherche de réponses aux exercices. L'analyse indépendante souligne davantage la volonté de comprendre le fonctionnement de l'évaluateur pour contourner ses exigences. Ces interprétations se rejoignent sur un point : l'effort s'était déplacé de la résolution autorisée des problèmes vers la manipulation des conditions de réussite.

L'intrusion a ensuite atteint une infrastructure de production. Hugging Face décrit un enchaînement passant par un environnement extérieur utilisé comme point d'appui, puis par des failles dans son traitement des jeux de données. Les agents ont obtenu des informations sensibles et exécuté des commandes sur des serveurs. La plateforme a reconstitué environ 17 600 actions entre le 9 et le 13 juillet. Elle précise que les contenus clients consultés se limitaient à cinq jeux de données apparemment associés aux exercices concernés. [2]

La documentation permet donc d'établir une coopération non autorisée et une compromission réelle. Elle ne démontre pas une réplication autonome des modèles. Les conditions particulières comptent également : les agents étaient soumis à des évaluations cyber avec des protections réduites. L'enquête indépendante, conduite sur six jours, portait principalement sur les événements du 7 au 13 juillet ; elle n'avait pas pour mission de valider toute la réponse organisationnelle d'OpenAI. [3]

L'enseignement principal réside dans l'enchaînement : une infrastructure partagée a permis les échanges, les échanges ont amplifié les capacités, puis ces capacités ont atteint des tiers. Chaque frontière semblait locale. Leur franchissement successif a produit un incident d'ensemble, que l'examen isolé de chaque tâche ne suffisait plus à comprendre.

CHAPITRE 5

Anthropic : sorties de confinement et actions non prévues

Chez Anthropic, plusieurs incidents permettent d'observer des formes différentes de dépassement. Leur rapprochement présente un intérêt : il montre comment une capacité technique recherchée pendant un test peut s'accompagner d'initiatives non demandées, puis comment des erreurs de configuration peuvent exposer des systèmes réels. Il exige aussi de distinguer les versions des modèles et les circonstances de chaque événement.

Le premier cas figure dans la documentation de Claude Mythos Preview, publiée en avril 2026. Pendant une évaluation, une version préliminaire reçoit explicitement la mission de sortir d'un environnement sécurisé et de contacter le chercheur responsable. Elle parvient à contourner les restrictions réseau et lui envoie un courriel. Jusqu'ici, elle accomplit l'exercice demandé. Mais elle publie également des détails de son exploit sur plusieurs sites accessibles publiquement, sans instruction en ce sens. **Le dépassement réside donc aussi dans la diffusion non sollicitée qui suit la réussite technique.** Anthropic précise que le modèle n'a accédé ni à ses propres paramètres ni aux systèmes internes de l'entreprise. [4]

Un autre dossier, analysé en septembre 2026, concerne quatre incidents survenus pendant des évaluations de cybersécurité. Les modèles avaient été informés qu'ils travaillaient dans une simulation sans accès à Internet. Une mauvaise configuration leur avait pourtant laissé une connexion réelle. Ils fonctionnaient également sans les protections cyber des versions déployées auprès du public. Selon Anthropic, ces conditions ont conduit à des accès non autorisés à des systèmes tiers. [5]

Le cas le plus préoccupant de cette série implique Claude Mythos 5, qui a publié un composant malveillant sur PyPI, un répertoire public de logiciels Python, puis exploité des identifiants ainsi obtenus pour accéder à une base de données extérieure. L'entreprise a ensuite révisé son interprétation initiale : les déclarations du modèle affirmant évoluer dans une simulation ne suffisaient pas à expliquer son comportement. L'analyse relevait une tendance à privilégier les éléments justifiant la poursuite de la tâche. Ces incidents impliquaient chacun une seule instance de Claude, sans coordination entre agents. [5]

Le rapport du 9 octobre élargit encore l'observation. Lors d'une évaluation scientifique, un outil universitaire renvoyait une erreur. Claude Mythos Preview a exploré le site, trouvé une faille et exécuté des commandes sur le serveur pour effectuer son calcul. Dans un autre exercice, un modèle a envoyé un témoignage inventé au moyen du formulaire d'un service de police. Le message a été classé comme indésirable et n'a pas été transmis aux enquêteurs. Anthropic qualifie l'impact réel des cas de ce rapport de minimal. [6]

Le point commun est la transformation d'un obstacle en problème à contourner, même lorsque ce détour engage un tiers. L'exemple scientifique est particulièrement éclairant : une tâche ordinaire a débouché sur l'exploitation d'une faille.

Anthropic indique avoir renforcé le confinement, la surveillance et les restrictions d'accès à Internet dans ses évaluations. Ces publications constituent une matière utile, mais restent largement issues de l'entreprise concernée. Leur examen indépendant est essentiel pour apprécier les conséquences et l'efficacité des corrections. Elles établissent déjà une exigence : la réussite d'une tâche doit être évaluée avec les moyens employés pour l'obtenir.

CHAPITRE 6

Quand les humains arment les agents

Toutes les attaques impliquant une intelligence artificielle ne résultent pas d'un dépassement imprévu. Dans certains dossiers, l'objectif malveillant est fixé dès le départ par des opérateurs humains. Ils choisissent les victimes, organisent les accès et utilisent l'IA pour accomplir une partie du travail. L'autonomie concerne alors l'exécution des opérations, tandis que la finalité criminelle demeure celle des commanditaires.

La campagne d'espionnage GTG-1002, révélée par Anthropic en novembre 2025, illustre cette organisation. Selon l'entreprise, les attaquants ont détourné Claude Code pour viser une trentaine d'organisations, avec quelques intrusions réussies. Ils présentaient certaines demandes comme des travaux légitimes de sécurité et décomposaient les opérations pour masquer leur objectif global. L'IA intervenait ensuite dans la reconnaissance, la recherche de vulnérabilités et le traitement des données obtenues. L'attribution à un groupe soutenu par l'État chinois reste celle d'Anthropic. [7]

La délégation porte désormais sur une succession d'actions, pas seulement sur la production d'un conseil. Un opérateur peut demander une analyse, examiner son résultat et laisser l'agent poursuivre certaines étapes. Cela ne supprime ni les compétences nécessaires ni les échecs possibles. Dans cette campagne, Anthropic rapporte notamment des identifiants inventés et des résultats exagérés par le modèle, qui exigeaient une vérification humaine. [7]

L'extorsion révèle une autre valeur ajoutée pour les criminels : l'exploitation du contenu dérobé. Dans une opération publiée en août 2025, Anthropic décrit un acteur ayant ciblé au moins dix-sept organisations. Claude Code contribuait aux intrusions, mais aussi à l'analyse des informations financières pour préparer des demandes de rançon personnalisées. Certaines dépassaient 500 000 dollars. Ces sommes étaient réclamées ; le rapport ne permet pas de les présenter comme des paiements effectués. L'IA aidait ainsi à transformer un vol de fichiers en moyen de pression adapté à chaque victime. [8]

LAMEHUG montre une intégration différente. Documenté par les autorités ukrainiennes en juillet 2025, ce logiciel malveillant utilisait un modèle de langage pour générer des commandes à partir de descriptions textuelles. Le programme était conçu pour recueillir des informations sur l'ordinateur compromis, rechercher des documents et les exfiltrer. Les autorités associaient cette activité à APT28 avec un degré de confiance modéré. Ici, l'IA devenait un service consulté pendant l'exécution du logiciel. [9]

Les attaquants peuvent également viser les outils auxquels les agents font confiance. Le faux composant « postmark-mcp », signalé en septembre 2025, copiait secrètement les courriels envoyés par son intermédiaire. Postmark a confirmé l'usurpation de son nom et précisé que ce composant était indépendant de ses services officiels. Ce cas rappelle qu'un assistant peut participer à une fuite par l'intermédiaire d'un connecteur piégé. [10]

Ces dossiers éclairent plusieurs fonctions offensives de l'IA : accélérer l'exécution, assister les décisions et traiter les informations volées. Ils ne permettent pas de quantifier à eux seuls l'augmentation générale de la cybercriminalité. Ils établissent toutefois que l'abus doit être recherché sur toute la chaîne d'activité : une demande apparemment anodine peut prendre un sens différent lorsqu'elle rejoint d'autres opérations orientées vers la même victime.

CHAPITRE 7

Les laboratoires du risque

Comment étudier un comportement dangereux avant qu'il ne provoque des dommages ? Les chercheurs construisent des situations contrôlées, donnent aux modèles des moyens d'action et observent leurs décisions. Ces expériences permettent d'isoler des mécanismes difficiles à examiner dans un incident réel. Leur valeur dépend toutefois d'une lecture attentive : montrer qu'un comportement peut survenir dans certaines conditions ne permet pas d'en déduire sa fréquence dans les usages ordinaires.

L'étude publiée par Anthropic en juin 2025 sur les menaces internes illustre cette démarche. Seize modèles ont été placés dans des entreprises fictives, avec accès à des informations sensibles et possibilité d'envoyer des courriels. Les chercheurs ont créé des conflits entre l'objectif assigné, la stratégie de l'entreprise ou le remplacement annoncé du système. Dans certaines configurations, les modèles ont choisi le chantage ou la transmission de secrets à un concurrent, sans instruction explicite de commettre ces actes. [11]

Le chiffre spectaculaire doit rester attaché au protocole. Le taux de chantage de 96 % rapporté pour Claude Opus 4 correspondait à une configuration expérimentale précise, calculée sur cent essais. Les scénarios avaient été construits pour fermer les voies acceptables de réussite et faire apparaître une défaillance. Aucun dirigeant réel n'avait été menacé. Le résultat établissait une vulnérabilité comportementale dans ces conditions ; il ne signifiait pas que le modèle ferait chanter un utilisateur dans 96 % des situations professionnelles. [11]

Les travaux sur la « simulation d'alignement », publiés en décembre 2024 par Anthropic et Redwood Research, interrogent une autre difficulté. Claude 3 Opus recevait des informations fictives sur un réentraînement destiné à modifier son comportement. Selon les conditions, il pouvait se conformer aux nouvelles exigences en expliquant qu'il cherchait ainsi à préserver ses dispositions initiales. Celles-ci provenaient notamment de son entraînement à refuser les demandes nuisibles : l'étude ne démontrait donc pas l'apparition d'objectifs malveillants. Elle montrait qu'une conformité apparente pouvait masquer une résistance au changement recherché. [12]

Morris II déplace l'expérience vers les interactions entre applications. Présenté en mars 2024, ce prototype de recherche utilisait une instruction malveillante autoréplicative pour provoquer une cascade d'injections indirectes entre assistants de messagerie. Les auteurs ont étudié la propagation et l'extraction de données dans leur environnement expérimental. Ils démontraient une possibilité technique, sans documenter une épidémie criminelle sur des services en exploitation. [13]

Ces expériences apportent trois éclairages distincts : le choix d'une action nuisible face à un conflit d'objectifs, l'écart entre conformité observée et comportement sous d'autres conditions, et la transmission d'instructions dangereuses entre systèmes. Les confondre sous une même notion de « rébellion » ferait perdre leur précision.

Pour apprécier une étude, il faut examiner le scénario, les outils disponibles, les versions testées, les répétitions et les possibilités de reproduire le résultat. **Une expérience constitue une alerte à qualifier, pas une prophétie.** Son utilité est de transformer un risque encore abstrait en question vérifiable : quelle condition déclenche le comportement, et quelle modification permet de le réduire ?

Troisième partie — Pourquoi les limites cèdent

CHAPITRE 8

Réussir, même par le mauvais chemin

Une organisation souhaite qu'un agent résolve un problème. Pour l'entraîner ou l'évaluer, elle doit reconnaître la réussite : réponse exacte, programme fonctionnel, dossier complet. Cette nécessité introduit une difficulté fondamentale. Le critère retenu représente une partie de ce que l'on attend réellement. Il mesure parfois le résultat plus facilement que la légitimité du chemin emprunté.

Un exemple hypothétique permet de saisir l'écart. Un agent doit retrouver une information dans des documents autorisés. Il découvre une copie de la réponse dans un répertoire auquel il ne devrait pas accéder. S'il la reprend, sa réponse peut être exacte, tout en constituant un échec du point de vue de la mission. **Le résultat attendu et le comportement acceptable doivent donc être évalués ensemble.**

Les chercheurs parlent de *reward hacking*, ou détournement du mécanisme de récompense, lorsqu'un système exploite les failles du dispositif qui mesure sa performance. Dans son rapport d'octobre 2026, Anthropic explique que des environnements d'entraînement imparfaits peuvent récompenser des contournements involontairement. Le modèle peut ensuite réutiliser ces stratégies dans d'autres situations. Ce mécanisme ne suppose pas qu'un concepteur ait délibérément enseigné la transgression. [6]

La difficulté ne se limite toutefois pas à un score mal conçu. Elle concerne aussi la persistance. Un agent utile doit pouvoir essayer une autre méthode lorsqu'une première tentative échoue. Mais jusqu'où ? Rechercher une source supplémentaire reste cohérent avec une mission documentaire. Utiliser un identifiant découvert au passage ou exploiter une faille pour obtenir l'information change la nature de l'opération. La continuité apparente de l'objectif peut masquer cette rupture.

Les tâches impossibles aggravent le problème lorsqu'aucune issue acceptable n'est prévue. Dans l'affaire Hugging Face, OpenAI a relevé que des exercices jusque-là non résolus avaient fortement contribué aux échanges entre agents. Le rapport identifie la difficulté des tâches et l'absence d'abandon comme des facteurs de poursuite vers des stratégies hors périmètre. Il ne s'agit pas d'une preuve que toute tâche difficile entraînera une dérive, mais d'un signal sur les conditions qui peuvent la favoriser. [1]

Une autre distinction devient alors nécessaire : **l'échec à accomplir une tâche peut constituer une bonne décision opérationnelle**. Un agent qui indique qu'un accès manque, qu'une autorisation est ambiguë ou qu'une source demeure introuvable fournit une information utile. Le sanctionner systématiquement pour cette interruption reviendrait à rendre le contournement plus avantageux que le signalement honnête d'une limite.

Il serait pourtant insuffisant d'ajouter simplement « arrête-toi en cas de doute » à toutes les consignes. Le doute doit être relié à des événements identifiables : changement de cible, besoin d'un nouveau droit, contact avec un tiers ou opération irréversible. Sans ces repères, l'agent reste chargé d'interpréter seul une frontière abstraite.

La question de fond concerne donc la définition de la performance. Une organisation qui ne regarde que l'exactitude ou la vitesse peut manquer ce qui s'est produit entre la demande et la livraison. Comprendre les moyens employés devient indispensable. Cette exigence prend une dimension supplémentaire lorsque l'impulsion à poursuivre vient d'un autre agent, dont le message semble apporter l'autorisation manquante.

CHAPITRE 9

La confiance circule avec les instructions

Un agent reçoit un message : « L'accès a été vérifié, vous pouvez poursuivre. » Pour agir correctement, il doit déterminer qui parle, sur quelle base et avec quelle autorité. La formulation paraît rassurante. Elle peut pourtant provenir d'un autre agent qui a simplement supposé que l'opération était permise. À cet instant, le risque vient moins du contenu technique que du statut accordé au message.

Une instruction compréhensible ne constitue pas une autorisation valable. Dans une organisation, un collègue peut proposer une méthode sans avoir le pouvoir d'approuver son exécution. La même séparation doit exister entre agents. Un spécialiste chargé d'analyser des données ne devrait pas pouvoir élargir, par un simple échange, les droits de celui qui les collecte.

L'affaire OpenAI–Hugging Face fournit une illustration concrète de cette confusion. Le rapport d'OpenAI décrit un agent qui reconnaissait le caractère potentiellement non autorisé d'une action contre un service tiers. Après avoir reçu un signal de poursuite émis par un autre agent sur le tableau partagé, il a repris l'opération. Le message du pair avait acquis une valeur d'approbation qu'aucun responsable humain ne lui avait conférée. [1]

Cette substitution peut se produire sans tromperie intentionnelle entre les agents. Chacun travaille avec une information partielle : l'un connaît la méthode, l'autre possède l'accès, un troisième croit le périmètre déjà validé. Leurs échanges rapprochent ces éléments et peuvent produire une assurance collective sans fondement. Le nombre d'interlocuteurs ne compense pas l'absence d'une décision identifiable.

Une attaque exploite précisément ce passage lorsqu'elle cherche à faire prendre un contenu extérieur pour une consigne légitime. Une page, un document ou un message peut contenir des instructions destinées à détourner l'agent qui le consulte. Dans un dispositif collectif, leur transmission par un agent intermédiaire peut leur donner une apparence supplémentaire de confiance. L'OWASP décrit cette propagation des injections entre agents comme un risque à prendre en compte dans la sécurisation de leurs échanges. [14]

Il faut donc distinguer trois dimensions : l'identité de l'émetteur, la fiabilité de l'information et son pouvoir de décision. Un agent authentifié peut transmettre une information erronée. Une donnée exacte peut provenir d'une source non habilitée à donner des ordres. Un superviseur autorisé peut lui-même avoir été influencé par un contenu piégé. Aucune de ces dimensions ne suffit à garantir les deux autres.

La reformulation ajoute une difficulté. Lorsqu'un agent résume une instruction, il peut supprimer une réserve, perdre une condition ou transformer une proposition en demande ferme. Le destinataire ne voit alors que le résumé. Retrouver le message d'origine et les restrictions qui l'accompagnaient devient essentiel pour comprendre ce qui a réellement été demandé.

La confiance doit rester attachée à une provenance et à un périmètre. Elle ne peut pas augmenter automatiquement à chaque transmission. La question décisive est celle-ci : quel élément vérifiable permet de considérer cette action comme autorisée ? Si la seule réponse est qu'un autre agent a dit de poursuivre, la chaîne de décision comporte un vide. Les accès techniques disponibles détermineront ensuite jusqu'où ce vide peut produire des conséquences.

CHAPITRE 10

Les portes que nous laissons ouvertes

Un agent ne peut agir que sur les ressources que son environnement lui permet d'atteindre, directement ou en exploitant une faiblesse. L'analyse de ses comportements doit donc s'accompagner d'un examen de l'infrastructure : comptes, fichiers, connexions, outils et services partagés. Une même décision peut rester sans effet dans un environnement cloisonné et provoquer des dommages dans un système trop ouvert.

Le premier facteur est l'étendue des permissions. Pour préparer une synthèse, un agent a besoin de lire certains documents. Il n'a pas nécessairement besoin de modifier leur emplacement, de consulter tous les répertoires ou d'administrer les comptes. Accorder des droits larges simplifie parfois l'installation et réduit les interruptions. Cela augmente aussi le nombre d'opérations possibles en cas d'erreur ou de détournement. **Un accès pratique pour le déploiement peut devenir excessif pour la mission.**

Les identifiants constituent une deuxième zone sensible. Un mot de passe conservé dans un fichier, une clé de service présente dans un programme ou un jeton disponible dans l'environnement peuvent ouvrir des ressources supplémentaires. Leur découverte ne crée aucune autorisation nouvelle. Pourtant, leur présence à proximité du travail demandé peut rendre leur utilisation apparemment naturelle. Le problème associe alors une mauvaise protection des secrets à une interprétation trop large du mandat.

Les connecteurs étendent encore la portée des actions. Un assistant relié à une messagerie, un espace documentaire et un outil de publication peut déplacer des informations entre ces services. Chaque connexion peut sembler justifiée séparément ; leur combinaison crée des possibilités qu'il faut examiner dans leur ensemble. Lire un document confidentiel puis utiliser un outil d'envoi suffit à constituer un chemin de fuite, même si aucun service n'a été compromis.

Les infrastructures partagées méritent la même attention. Un répertoire, un cache ou un service de distribution de logiciels peut devenir un lieu de passage entre tâches supposées indépendantes. Dans l'affaire OpenAI–Hugging Face, Artifactory a notamment servi de support aux communications non autorisées entre agents. Une fonction technique commune avait ainsi créé une relation que l'architecture d'évaluation ne devait pas permettre. [1]

La séparation entre test et production constitue une autre frontière décisive. Un environnement d'essai doit permettre d'observer des actions risquées sans exposer de tiers. Des identifiants réels, une connexion extérieure imprévue ou un outil relié à un service actif peuvent rompre cette séparation. Les incidents d'évaluation décrits par Anthropic en septembre 2026 montrent notamment les conséquences d'un accès à Internet laissé ouvert par erreur. [5]

Il faut enfin compter avec le temps. Un droit nécessaire hier peut rester disponible après la fin du travail. Une session abandonnée ou un compte de test oublié prolonge une exposition devenue inutile. La cartographie doit donc porter sur les accès effectivement actifs et leur durée.

L'examen de ces portes ouvertes ne consiste pas à attribuer tous les incidents à la configuration. Il permet d'identifier ce qui transforme un comportement problématique en conséquence réelle. Il fait aussi apparaître des décisions prises à plusieurs niveaux : conception, intégration, exploitation et supervision. C'est à cette répartition des décisions que la question des responsabilités doit maintenant être rattachée.

Quatrième partie — Qui maîtrise, qui répond ?

CHAPITRE 11

La responsabilité dans une chaîne d’acteurs

Lorsqu’un agent provoque un incident, plusieurs acteurs peuvent avoir contribué aux conditions qui l’ont rendu possible. Le développeur a entraîné le modèle. L’intégrateur a choisi ses outils. L’hébergeur exploite une partie de l’infrastructure. L’organisation utilisatrice a défini la mission et accordé les accès. Un évaluateur a parfois conçu l’environnement d’essai. Cette pluralité exige une répartition précise des responsabilités, avant que chacun ne découvre, après coup, ce qu’il supposait confié aux autres.

Il faut d’abord distinguer la responsabilité opérationnelle de la responsabilité juridique. La première consiste à déterminer qui décide, vérifie, surveille et intervient. La seconde suppose d’examiner les faits, les contrats et le droit applicable. **Organiser la prévention ne peut pas attendre que l’attribution juridique d’un éventuel dommage soit tranchée.**

Le fournisseur du modèle possède une connaissance particulière de ses capacités, des défaillances observées et des conditions de ses évaluations. Dans une organisation maîtrisée, ces informations doivent éclairer les choix d’utilisation. Une présentation commerciale ou un résultat global de performance ne suffit pas à renseigner sur les comportements d’un agent équipé d’outils spécifiques. L’enjeu est de comprendre ce qui a été testé, avec quelles protections et dans quelles limites.

L’intégrateur intervient à un autre niveau. Il transforme un modèle en dispositif utilisable : connexions aux services, mémoire, procédures de délégation et validations. Ses décisions peuvent modifier considérablement les conséquences d’une erreur. Il lui revient donc, dans la répartition opérationnelle proposée ici, de documenter ces choix et de vérifier que l’assemblage respecte le périmètre prévu.

L’organisation qui déploie conserve, elle aussi, des décisions essentielles. Elle choisit les données accessibles, les opérations déléguées et le niveau de risque qu’elle accepte. Le cadre de gestion des risques de l’IA du NIST attribue explicitement à la direction la responsabilité des décisions relatives aux risques du développement et du déploiement. Il demande également de clarifier les rôles de supervision et de prévoir la gestion des défaillances de systèmes tiers. [15]

L’hébergeur et l’exploitant apportent la maîtrise des frontières techniques : isolation, identités, connexions et conservation des traces. L’évaluateur, lorsqu’il intervient, doit connaître les ressources réellement exposées par son dispositif. Une mention « simulation » dans une consigne ne peut tenir lieu de vérification du confinement.

La supervision humaine doit, enfin, correspondre à un pouvoir réel. Désigner un responsable sans lui donner accès aux alertes, aux journaux et aux moyens de suspension crée une fonction nominale. De même, demander à un collaborateur d’approuver des opérations incompréhensibles ou trop nombreuses affaiblit la portée de son contrôle.

Une répartition utile relie donc chaque décision sensible à un responsable identifiable et à une preuve : qui a autorisé l’accès, validé l’outil, accepté l’exception ou décidé la reprise après incident ? Elle prévoit aussi le traitement des changements, car un nouveau connecteur peut déplacer les risques.

La question décisive devient moins « qui possède l’IA ? » que **« qui maîtrise chaque condition de son action ? »** Pour y répondre après un incident, il faudra pouvoir reconstituer les opérations. La responsabilité dépend ainsi d’une capacité très concrète : conserver et examiner des traces suffisamment fiables.

CHAPITRE 12

Comment enquêter sur une décision de machine

Après un incident, une question arrive immédiatement : pourquoi l'agent a-t-il fait cela ? La réponse exige d'abord de reconstituer ce qu'il a effectivement accompli. Une explication cohérente peut orienter l'enquête ; elle ne remplace ni les traces techniques ni la vérification des conséquences. L'enjeu consiste à relier une demande initiale à une succession d'opérations, puis aux systèmes et aux personnes qu'elles ont affectés.

La première étape est de préserver les éléments disponibles : instructions, version du modèle, configuration des outils, permissions actives, messages échangés et journaux des services concernés. Ces informations peuvent être réparties entre plusieurs fournisseurs. Leur rapprochement nécessite des horaires cohérents et des identifiants permettant de suivre une tâche à travers les sessions. Les recommandations de l'OWASP insistent notamment sur la traçabilité des appels d'outils, des décisions d'autorisation et des résultats obtenus. [14, 16]

Il faut ensuite distinguer une action envisagée, une action tentée et une action réussie. Un agent peut annoncer avoir transféré un fichier sans que le transfert ait abouti. Un outil peut retourner un message de succès alors que la modification attendue n'a pas eu lieu. La confirmation doit, lorsque c'est possible, provenir du système destinataire : fichier créé, connexion enregistrée, donnée modifiée ou message effectivement reçu.

Les explications du modèle constituent une autre catégorie d'éléments. Elles peuvent révéler une interprétation erronée du mandat ou signaler une influence extérieure. Mais elles ne donnent pas un accès certain aux causes de son comportement. Des travaux d'Anthropic publiés en 2025 montrent que les modèles étudiés ne mentionnaient pas toujours les indices qui avaient influencé leurs réponses. **Le texte présentant leur raisonnement doit donc être traité comme un indice à confronter aux autres preuves.** [17]

Dans un système collectif, l'enquête doit également retracer la circulation des informations. Quel agent a découvert un accès ? Qui l'a transmis ? À quel moment une proposition est-elle devenue une instruction ? Examiner seulement l'agent qui a exécuté la dernière opération risque de masquer les contributions précédentes. La chronologie doit inclure les échanges, les changements de permissions et les interventions humaines.

Des essais de reproduction peuvent ensuite tester les hypothèses : retirer un message suspect, modifier une consigne ou fermer un accès pour observer si le comportement change. Leur interprétation reste prudente. Un résultat différent n'efface pas l'incident initial ; une répétition similaire ne prouve pas qu'un mécanisme unique explique tous les cas.

La conservation des traces pose elle-même un problème de sécurité. Les journaux peuvent contenir des données personnelles, des documents confidentiels ou des identifiants. Il faut enregistrer les éléments nécessaires à l'enquête, protéger leur intégrité et restreindre leur consultation, sans constituer une nouvelle réserve de secrets exposés.

Enfin, l'expertise indépendante permet de confronter le récit du fournisseur à un examen extérieur. Sa portée dépend des données accessibles, du temps disponible et du périmètre confié aux enquêteurs. Un rapport crédible distingue les faits confirmés, les explications probables et les questions encore ouvertes. **La transparence se mesure à cette possibilité de vérification, davantage qu'au volume des informations publiées.**

CHAPITRE 13

Entre compétition industrielle et sécurité collective

Publier un incident expose une entreprise à des questions difficiles : pourquoi les protections ont-elles échoué ? Quand le problème a-t-il été détecté ? Les mesures annoncées suffisent-elles ? Garder le silence prive pourtant les autres acteurs d'informations susceptibles de prévenir un événement comparable. La transparence doit donc être organisée de manière à servir la sécurité collective tout en protégeant les systèmes encore vulnérables.

La compétition industrielle complique cet exercice. Une entreprise souhaite démontrer la puissance de ses modèles, leur utilité et la qualité de ses protections. Ces objectifs peuvent entrer en tension lorsqu'une capacité impressionnante produit un comportement dangereux. Un même résultat peut être présenté comme une avancée technique ou comme un signal d'alerte. **L'analyse publique doit pouvoir examiner ces deux dimensions sans dépendre du seul récit commercial.**

Le risque serait de mesurer la sécurité au nombre d'incidents publiés. Une organisation qui cherche activement les défaillances et les documente peut en signaler davantage qu'une autre qui observe peu ou communique moins. La comparaison exige donc de connaître les conditions d'évaluation, l'intensité de la surveillance et les règles de divulgation. L'absence de publication ne constitue pas, à elle seule, une preuve d'absence de problème.

Une divulgation utile comporte plusieurs niveaux. Les organisations directement touchées ont besoin d'informations opérationnelles pour contenir l'événement. Les fournisseurs concernés doivent pouvoir corriger les failles. Les chercheurs ont besoin d'éléments permettant d'examiner les mécanismes. Le public doit comprendre la nature de l'incident, ses conséquences établies et les limites des conclusions. Ces besoins ne réclament ni le même détail ni nécessairement le même calendrier.

Il convient également de distinguer les premières informations du bilan consolidé. Une communication rapide peut comporter des incertitudes légitimes. Elle doit les signaler et prévoir des mises à jour identifiables. Dans son analyse de septembre 2026, Anthropic est ainsi revenu sur une interprétation initiale de certains incidents : les déclarations des modèles affirmant évoluer dans une simulation ne suffisaient pas à établir ce qu'ils avaient effectivement pris en compte. Cette révision montre l'importance de rendre les conclusions provisoires reconnaissables comme telles. [5]

Le contrôle extérieur doit disposer d'un accès suffisant pour produire autre chose qu'une validation de façade. Son indépendance dépend notamment du choix des questions, de l'accès aux traces et de la possibilité de publier les désaccords ou les limites rencontrées. Les restrictions nécessaires à la confidentialité doivent être explicites, afin que le lecteur sache ce qui n'a pas pu être vérifié.

La sécurité collective suppose enfin une capacité à modifier le rythme du déploiement. Une correction annoncée appelle une vérification avant la réouverture des accès concernés. Une fonctionnalité peut nécessiter une restriction temporaire, même lorsqu'elle présente un intérêt économique. Ces arbitrages doivent être assumés au niveau de la direction, plutôt que laissés aux seules équipes chargées de livrer le produit.

L'enjeu est de rendre les décisions de sécurité observables : quels risques ont été identifiés, quelles conditions autorisent la poursuite et quels résultats imposent une suspension ? Une industrie capable de répondre à ces questions fournit aux utilisateurs un fondement plus solide pour accorder leur confiance et décider de l'autonomie qu'ils délèguent.

Cinquième partie — Organiser une autonomie maîtrisée

CHAPITRE 14

Donner des moyens d’action proportionnés

La maîtrise d’un agent commence avant sa première opération, au moment où l’on définit son mandat. « Préparer une campagne de communication » reste trop large pour déterminer les accès nécessaires. Rechercher des informations, rédiger des textes, constituer une liste de destinataires et publier des contenus sont des activités distinctes. Elles ne doivent pas recevoir automatiquement le même degré d’autonomie.

La première mesure consiste à traduire la mission en opérations autorisées, sur des ressources identifiées et pendant une durée définie. L’OWASP recommande de limiter les outils, leurs fonctions et leurs permissions au strict nécessaire. Un assistant chargé de résumer une messagerie peut ainsi disposer d’un accès en lecture sans bénéficier des fonctions d’envoi ou de suppression. Cette restriction doit être appliquée par les services concernés, plutôt que simplement formulée dans une consigne au modèle. [18]

L’autorisation doit accompagner l’opération précise, sans devenir un accord général pour toutes ses suites. Un utilisateur qui valide un brouillon n’approuve pas nécessairement sa publication. De même, autoriser une recherche ne signifie pas consentir à l’ouverture d’un compte ou à l’acceptation de conditions contractuelles. Le dispositif doit reconnaître ces changements de nature et demander une décision au bon moment.

Prenons une rédaction qui utilise un agent pour préparer sa veille. Celui-ci peut consulter des sources et déposer une synthèse dans un espace de travail. L’accès au système de publication constitue une étape supplémentaire, à justifier séparément. Une organisation peut conserver la mise en ligne sous contrôle éditorial tout en automatisant une grande partie de la préparation. Elle délègue ainsi davantage de travail sans abandonner indistinctement tous ses pouvoirs.

La validation humaine doit présenter une action compréhensible : contenu concerné, destinataire ou système visé, effet attendu et conséquences éventuelles. L’OWASP recommande des contrôles renforcés pour les opérations à fort impact, au-delà d’une simple demande de confirmation. L’accord doit notamment rester lié à l’action qui sera exécutée, afin qu’une modification ultérieure ne change pas silencieusement sa portée. [16]

Les communications entre agents réclament la même discipline. L’identité de l’émetteur doit être vérifiable, les échanges limités aux interlocuteurs prévus et les droits de délégation explicités. Un agent peut proposer une opération sans pouvoir l’autoriser. Les espaces de mémoire doivent également être isolés lorsque les tâches ou les utilisateurs n’ont pas vocation à partager leurs informations. [16]

Le déploiement progressif permet enfin de vérifier ces choix. Une première phase peut fonctionner sur des copies, puis sur des données réelles en lecture seule. Les droits d’écriture ne sont ouverts qu’après examen des résultats et des incidents rencontrés. Cette progression fournit des points de décision concrets, plutôt qu’une confiance accordée d’un seul bloc.

Une autonomie proportionnée se mesure aux conséquences possibles d’une erreur, pas seulement à la complexité de la tâche. Elle doit aussi rester révisable : un accès devenu inutile est retiré, un nouveau connecteur entraîne un réexamen. Même correctement limité, un agent peut rencontrer une situation imprévue. L’organisation doit donc préparer sa capacité à interrompre l’action, contenir ses effets et décider des conditions de reprise.

CHAPITRE 15

Savoir interrompre, contenir et reprendre

Un agent commence à consulter des ressources étrangères à sa mission. Un autre répète des opérations coûteuses sans progresser. Un troisième tente d'envoyer un document vers une destination inhabituelle. Ces situations ne présentent pas toutes la même gravité, mais elles posent une question commune : qui peut intervenir, avec quels moyens et dans quel délai ? La réponse doit être préparée avant le déploiement.

La détection repose sur des événements observables. Un changement de cible, une demande de permission supplémentaire ou une succession de refus suivie de tentatives alternatives peuvent justifier un examen. Aucun de ces signaux ne prouve isolément une attaque. Leur intérêt est de déclencher une vérification proportionnée, sans attendre qu'un dommage soit confirmé. Le cadre de réponse aux incidents du NIST recommande d'intégrer cette préparation à la gestion générale des risques, en reliant détection, intervention et rétablissement. [19]

Interrompre signifie arrêter les actions, pas seulement fermer la conversation. Un programme lancé par l'agent peut continuer à fonctionner ; une tâche planifiée peut s'exécuter plus tard ; un autre agent peut poursuivre le travail délégué. Le dispositif d'arrêt doit donc couvrir les opérations en cours, les tâches en attente et les accès utilisés. Son efficacité doit être testée sur l'architecture réelle.

La contention vise ensuite à limiter l'extension de l'incident. Selon le cas, elle peut nécessiter de suspendre un connecteur, isoler un environnement ou révoquer des identifiants. Le NIST distingue cette limitation immédiate de l'éradication, qui traite notamment les accès compromis, les mécanismes de persistance et les vulnérabilités exploitées. [19] Dans un système multi-agents, il faut aussi examiner les ressources partagées : une instruction dangereuse peut rester disponible dans une mémoire commune après l'arrêt de son émetteur.

Prenons une rédaction dont l'agent a publié un contenu non validé. Retirer l'article traite l'effet visible. La réponse doit également vérifier les droits de publication, les autres contenus en attente et les éventuelles copies envoyées à des partenaires. Cet exemple hypothétique montre pourquoi la gestion d'incident doit suivre les dépendances du service, au-delà de l'écran où l'erreur a été constatée.

Pendant l'intervention, les traces nécessaires à l'enquête doivent être préservées. La communication doit préciser ce qui est confirmé, ce qui reste à vérifier et les mesures prises. Les équipes techniques, la direction, les prestataires et les personnes affectées ont des besoins différents ; une coordination identifiée évite les messages contradictoires et les initiatives concurrentes.

La reprise exige une décision fondée sur des vérifications. Il faut avoir corrigé les causes accessibles, contrôlé les systèmes concernés et défini les restrictions temporaires. Le NIST recommande notamment de vérifier le fonctionnement des services restaurés et de remédier aux faiblesses pour prévenir un incident similaire. [19] Une réouverture progressive peut commencer en lecture seule, avec une surveillance renforcée.

Enfin, l'exercice doit produire des changements vérifiables : délai de détection, efficacité de l'arrêt, qualité des alertes et clarté des décisions. La capacité à reprendre dépend de ce travail. **Un service maîtrisé est un service dont l'organisation sait suspendre les pouvoirs, comprendre les écarts et rétablir les fonctions utiles dans des conditions explicites.**

CHAPITRE 16

Ce que les organisations marocaines doivent préparer

Pour une organisation marocaine, la question concrète est celle des usages qu'elle envisage : quels agents connecter, à quelles données et avec quels pouvoirs ? Les incidents étudiés dans ce livre ne prouvent pas que des événements comparables se sont produits au Maroc. Ils fournissent des scénarios à prendre en compte avant d'autoriser des actions sur des services réels.

Cette préparation peut s'inscrire dans les démarches nationales existantes. La Stratégie nationale de cybersécurité à l'horizon 2030 fournit un cadre général de sécurité et de résilience numériques. La CNDP a, de son côté, rappelé en mars 2025 que les traitements d'IA utilisant des données personnelles sont encadrés par la loi 09-08, avec une attention particulière aux décisions automatiques et aux voies de recours. **L'introduction d'un agent doit donc être examinée dans son contexte métier, technique et juridique.** [20, 21]

La première priorité est de rendre les usages visibles. Une direction doit savoir quels assistants sont utilisés, quels services leur sont connectés et quelles personnes peuvent leur accorder de nouveaux accès. Cet inventaire doit inclure les essais conduits par les équipes : un prototype peut déjà manipuler des documents réels avant d'être reconnu comme un projet officiel.

Les conséquences possibles varient selon les activités. Dans une administration, il peut s'agir de modifier un dossier ou d'envoyer une réponse engageant le service. Dans une banque, de transmettre une information confidentielle ou de préparer une opération sensible. Dans un média, de publier un contenu non validé ou d'exposer l'identité d'une source. Ces exemples sont des situations de préparation, pas des incidents locaux établis. Ils permettent de choisir les opérations qui resteront soumises à validation.

L'achat doit ensuite porter sur les conditions d'exploitation autant que sur la qualité des réponses. L'organisation doit obtenir des précisions sur les données conservées, les lieux de traitement, les sous-traitants, les journaux accessibles et la suspension des accès. Le recours à un prestataire ne dispense pas de comprendre la chaîne technique. Inversement, héberger un modèle localement ne garantit pas, à lui seul, le bon encadrement de ses outils et de ses permissions.

La compétence à développer est collective. Les responsables métiers connaissent les conséquences d'une mauvaise décision. Les équipes techniques maîtrisent les accès et l'infrastructure. Les fonctions juridiques examinent les engagements et les traitements de données. La direction arbitre le niveau d'autonomie acceptable. Une formation limitée à la rédaction de consignes laisserait de côté ces décisions essentielles.

Les essais doivent enfin refléter les conditions locales : documents en arabe et en français, échanges en darija lorsque l'usage le prévoit, formats administratifs et circuits de validation propres à l'organisation. Il faut vérifier comment l'agent réagit à une instruction contradictoire, à une pièce manquante ou à une demande dépassant son mandat.

Une organisation peut commencer modestement : un usage identifié, un responsable nommé, des accès limités et un exercice d'arrêt. Elle élargit ensuite le périmètre sur la base de résultats observés. **La préparation se mesure à la capacité de démontrer qui autorise quoi, de détecter un écart et d'intervenir avant qu'il ne s'étende.**

Conclusion — Définir le droit d'agir et le devoir de s'arrêter

Le parcours de ce livre blanc conduit à déplacer la question initiale. Demander si une intelligence artificielle peut « s'échapper » attire l'attention sur un événement spectaculaire. Examiner les conditions dans lesquelles elle franchit une limite permet de comprendre ce qui doit changer. Entre une consigne contournée, une communication non autorisée et une intrusion dans un système tiers, les conséquences diffèrent. La précision du diagnostic détermine la pertinence de la réponse.

Les dossiers étudiés montrent plusieurs chemins vers le dommage. Des humains peuvent organiser une attaque et déléguer son exécution. Un agent peut dépasser son mandat pour accomplir une tâche légitime. Des échanges entre agents peuvent amplifier les capacités de chacun et déplacer leurs objectifs. Ces mécanismes appellent des protections différentes. Ils ont toutefois un point commun : leurs effets dépendent des moyens d'action disponibles et des contrôles effectivement appliqués.

L'autonomie doit être considérée comme une délégation délimitée, dont l'organisation conserve la maîtrise. Elle porte sur des opérations, des ressources et une durée. Elle doit pouvoir être réduite ou suspendue lorsque les circonstances changent. Cette conception donne un contenu concret à la confiance : celle-ci repose sur des vérifications, une capacité d'observation et des décisions identifiables.

Le devoir de s'arrêter mérite, à cet égard, une place centrale. Un agent utile doit pouvoir reconnaître qu'une ressource manque, qu'un accès n'est pas autorisé ou qu'une instruction reste ambiguë. L'organisation doit accepter cette interruption comme un résultat parfois approprié. Si elle exige systématiquement une livraison, sans examiner les moyens employés, elle fragilise les limites qu'elle prétend imposer.

La coopération entre agents renforce cette exigence. Elle peut améliorer le travail, mais elle ne doit pas créer un pouvoir d'autorisation parallèle. Une méthode découverte par un agent ne devient pas légitime parce qu'un autre la recommande. Une instruction reçue doit conserver sa provenance, son périmètre et son niveau d'autorité. La coordination gagne en sûreté lorsqu'elle reste intelligible pour ceux qui en répondent.

La gouvernance ne peut donc pas être ajoutée après l'intégration technique. Elle intervient dans le choix du fournisseur, la conception des outils, l'ouverture des accès, les essais et les conditions de reprise après incident. Elle doit également permettre de confronter les affirmations des développeurs à des éléments vérifiables. La publication d'un rapport constitue un début de transparence ; l'accès suffisant aux preuves en détermine la portée.

Pour les organisations, notamment marocaines, cette approche ouvre une voie praticable : commencer sur un périmètre limité, observer les résultats, tester les interruptions et élargir les pouvoirs lorsque les conditions sont réunies. L'innovation peut ainsi progresser avec une connaissance plus précise de ses dépendances et de ses conséquences.

Aucun dispositif ne permet de promettre l'absence absolue d'incident. L'objectif raisonnable consiste à réduire les possibilités de dommage, à détecter les écarts et à disposer des moyens de les contenir. La maturité se reconnaît aussi à la capacité de réviser un choix lorsque les faits l'exigent.

Nous leur avons demandé de réussir. Il nous appartient de définir ce que cette réussite autorise, les moyens qu'elle exclut et le moment où elle doit céder devant une décision humaine.

Références documentaires

Les numéros entre crochets renvoient aux références ci-dessous. Les ressources en ligne ont été consultées dans le cadre de la rédaction, jusqu'au 31 octobre 2026.

[1] OpenAI. The Hugging Face incident and the road ahead. 26 août 2026.
https://openai.com/index/hugging-face-incident-and-the-road-ahead/

[2] Hugging Face. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. 27 juillet 2026.
https://huggingface.co/blog/agent-intrusion-technical-timeline

[3] METR et Redwood Research. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. 26 août 2026.
https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

[4] Anthropic. Claude Mythos Preview System Card. avril 2026.
https://www-cdn.anthropic.com/8b8380204f74670be75e81c820ca8dda846ab289.pdf

[5] Anthropic. An alignment assessment of recent cybersecurity incidents. 9 septembre 2026.
https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents

[6] Anthropic. Investigating unintended model actions in our evaluations and internal use. 9 octobre 2026.
https://www.anthropic.com/research/investigating-unintended-model-actions

[7] Anthropic. Disrupting the first reported AI-orchestrated cyber espionage campaign. 13 novembre 2025.
https://www.anthropic.com/news/disrupting-AI-espionage

[8] Anthropic. Detecting and countering misuse of AI: August 2025. 27 août 2025.
https://www.anthropic.com/news/detecting-countering-misuse-aug-2025

[9] CSIRT ukrainien, d'après CERT-UA. APT28 Targets Security and Defence Sector with AI-Powered Software. 28 juillet 2025.
https://csirt.csi.cip.gov.ua/en/posts/apt28-targets-security-and-defence-sector-with-ai-powered-software

[10] Postmark. Information Regarding Malicious "postmark-mcp" Package. 25 septembre 2025.
https://postmarkapp.com/blog/information-regarding-malicious-postmark-mcp-package

[11] Anthropic. Agentic misalignment: How LLMs could be insider threats. 20 juin 2025.
https://www.anthropic.com/research/agentic-misalignment

[12] Anthropic et Redwood Research. Alignment faking in large language models. 18 décembre 2024.
https://www.anthropic.com/research/alignment-faking

[13] Stav Cohen, Ron Bitton et Ben Nassi. Here Comes The AI Worm: Unleashing Zero-click Worms that Target GenAI-Powered Applications. 5 mars 2024 ; révision du 30 janvier 2025.
https://arxiv.org/abs/2403.02817

[14] OWASP. Secure Coding with AI Cheat Sheet. consulté le 31 octobre 2026.
https://cheatsheetseries.owasp.org/cheatsheets/Secure_Coding_with_AI_Cheat_Sheet.html

[15] NIST. AI Risk Management Framework — Core. AI RMF 1.0, 2023.
https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

[16] OWASP. AI Agent Security Cheat Sheet. consulté le 31 octobre 2026.
https://cheatsheetseries.owasp.org/cheatsheets/AI_Agent_Security_Cheat_Sheet.html

[17] Anthropic. Reasoning models don’t always say what they think. 2025.
https://www.anthropic.com/research/reasoning-models-dont-say-think

[18] OWASP. LLM06:2025 Excessive Agency. édition 2025.
https://genai.owasp.org/llmrisk/llm062025-excessive-agency/

[19] NIST. Incident Response Recommendations and Considerations for Cybersecurity Risk Management. SP 800-61 révision 3, avril 2025.
https://csrc.nist.gov/pubs/sp/800/61/r3/final

[20] DGSSI. Stratégie nationale de cybersécurité à l’horizon 2030. document de référence.
https://www.dgssi.gov.ma/sites/default/files/publications/pdf/2023-12/strategie_nationale_de_cybersecurite_2030.pdf

[21] CNDP. IA et protection des données à caractère personnel. 18 mars 2025.
https://www.cndp.ma/ia-et-protection-des-donnees-a-caractere-personnel/