

GUIDE 08

Claude prépare, l'humain engage sa responsabilité

Niveau 3 — choisir un niveau de contrôle humain proportionné aux conséquences d'une erreur, plutôt qu'uniforme.



Le contrôle nécessaire dépend moins de la tâche que de ce qui se passe si Claude se trompe.

Emplacement réservé à l'infographie de la collection (format portrait). Le schéma ci-dessus tient lieu de repère de mise en page.

Assistance n'est pas délégation

Claude est très utile pour synthétiser, explorer, comparer, structurer, rédiger et proposer des options. Ces capacités ne signifient pas que son résultat s'utilise de la même manière dans tous les contextes.

Faire préparer un travail ne transfère pas la responsabilité de ce travail. Ce qui change d'un cas à l'autre, ce n'est pas la difficulté de la tâche : c'est ce qui se produit si le résultat est faux et que personne ne s'en aperçoit.

Trois demandes, trois régimes

- 1 Conséquence faible**
« Propose trois titres pour ce document interne. » L'erreur est visible immédiatement et coûte peu.
- 2 Conséquence moyenne**
« Prépare une réponse à un client à partir du dossier. » Le texte peut créer un engagement ou un malentendu : relecture nécessaire.
- 3 Conséquence forte**
« Recommande une décision sur une personne, une obligation contractuelle ou un risque financier. » Le contrôle doit être fort — et parfois compétent au sens métier.

Ce que ce guide n'est pas. Il ne fournit pas d'avis juridique et ne qualifie aucun usage de Satellitem au regard d'une réglementation. Il propose un modèle de travail proportionné au risque, à confronter à la politique interne.

À RETENIR

Claude peut faire une partie du travail. Cela ne transfère pas automatiquement la responsabilité de ce travail.

VERSION WEB PERMANENTE

consulting.amplify-belgium.com/ressources/claude-prepare-lhumain-engage

GUIDE 08 · COLLECTION KNOWLEDGE CLAUDE

Vérifié le 28 août 2026 · 6 pages

GUIDE 08 · 01

Construire une échelle de contrôle

Un modèle mental simple, pas un calcul : **risque** $\approx$ **conséquence** × **difficulté de détection** × **difficulté de retour en arrière**. Les trois facteurs se multiplient — une erreur peu grave mais invisible et irréversible mérite plus d'attention qu'une grosse erreur immédiatement visible.

NIVEAU	CONDITIONS	EXEMPLES · CE QUE FAIT L'HUMAIN
1. Assistance libre	Impact faible, erreur visible, action réversible, aucun engagement externe.	Idées, reformulation, mise en structure. L'humain lit, prend ce qui l'intéresse, jette le reste.
2. Relecture ordinaire	Document professionnel, information interne, conséquence modérée.	Compte rendu, synthèse, e-mail de travail. L'humain relit les points importants avant diffusion.
3. Validation compétente	Information critique, interprétation métier, impact externe.	Données financières, engagement client, communication sensible. La relecture demande la compétence du domaine, pas seulement de l'attention.
4. Décision réservée	Droits d'une personne, sécurité, obligations juridiques, contexte à fort risque, décision difficilement réversible.	La décision reste humaine et suit un processus formel. Claude prépare la matière ; il ne conclut pas.

« Human in the loop » n'est pas un badge

Mettre un humain quelque part dans la chaîne ne dit rien du contrôle exercé. Ce qui compte, c'est *où* il intervient, avec *quelle* compétence et sur *quoi* il peut réellement revenir.

Le NIST recommande de définir les rôles humains et les processus de supervision en fonction du contexte : la gouvernance ne consiste pas à mettre un humain partout, mais à savoir où son intervention apporte une maîtrise réelle du risque. ^{[3][4]}

Le facteur le plus sous-estimé

La **difficulté de détection**. Une erreur spectaculaire se voit ; une date fausse dans un tableau de vingt lignes, non. C'est pourquoi le niveau se décide sur la nature de l'information autant que sur son enjeu apparent.

Corollaire pratique : plus une sortie est fluide et bien structurée, plus la détection est difficile — ce qui remonte le niveau exigé, pas l'inverse.

À RETENIR

Le niveau de contrôle se déduit des conséquences d'une erreur, pas de l'impression de difficulté de la tâche.

GUIDE 08 · 02

La validation humaine doit être réelle

Une personne qui clique «OK» sans vérifier n'est pas un contrôle. Elle est une trace de contrôle — ce qui est très différent, y compris en cas de problème.

Le biais d'automatisation

Lorsqu'un système produit une réponse claire et assurée, l'humain a tendance à lui accorder trop de confiance. Le phénomène est documenté et il est explicitement nommé dans le règlement européen sur l'IA, qui demande que les personnes chargées du contrôle restent conscientes de la tendance à se fier automatiquement ou excessivement à la sortie du système — l'*automation bias*.^[5]

Précision nécessaire. Cette obligation figure à l'article 14 du règlement, qui s'applique aux systèmes d'IA classés à **haut risque**. Elle ne constitue pas une règle juridique applicable à chaque conversation avec Claude. Le phénomène décrit, lui, est parfaitement pertinent au quotidien.^[5]

VALIDATION DE FAÇADE

«Ça a l'air correct, j'envoie.»

Le jugement porte sur l'apparence du texte. Aucun élément extérieur n'a été consulté — donc rien n'a été vérifié, au sens strict.

Cinq conditions d'une validation utile

Compétence — la personne comprend le sujet.

Information — elle dispose des sources nécessaires pour vérifier.

Temps — la relecture n'est pas symbolique.

Autorité — elle peut réellement refuser ou corriger.

Traçabilité, lorsque c'est nécessaire — on sait qui a validé quoi.

Retirer une seule de ces conditions transforme la validation en formalité.

VALIDATION RÉELLE

«Je contrôle les trois chiffres critiques dans la source, je vérifie l'engagement commercial, puis j'envoie.»

Trois points nommés à l'avance, une source consultée, une décision explicite. C'est reproductible et défendable.

Le lien avec le Guide 04. Le Guide 04 explique *comment* vérifier une réponse. Ce guide répond à une autre question : *qui* doit la vérifier, et avec quel niveau d'exigence. Les deux sont nécessaires ; aucun ne remplace l'autre.

À RETENIR

Une validation n'est utile que si le valideur peut comprendre, contester et arrêter.

GUIDE 08 · 03

Préparer, recommander, décider, agir

Quatre degrés d'implication, souvent confondus dans une même phrase — « Claude s'occupe de ça ». Les distinguer permet de savoir ce qui reste à faire, et par qui.

DEGRÉ	CE QUE FAIT CLAUDE	CE QUE DOIT FAIRE L'HUMAIN
1. Préparer	Collecte, résumé, reformule, met en forme, propose des options.	Contrôle normal : relire les points qui comptent.
2. Recommander	Compare des options et suggère une voie.	Vérifier les critères retenus, les sources et les angles morts possibles. Une recommandation reflète les critères qu'on lui a donnés.
3. Décider	Le résultat détermine une décision à conséquence réelle.	Humain compétent, politique claire, parfois processus formel. La décision n'est pas déléguée.
4. Agir	Exécute une action externe via un outil connecté.	Combiner la responsabilité décrite ici et les permissions du Guide 07 : ce qui est autorisé, ce qui exige une approbation, ce qui est bloqué. ^[1]

Quelques usages courants, situés

USAGE	CLAUDE PEUT	L'HUMAIN DOIT
Synthèse de réunion	Préparer	Relire les décisions et les échéances
Proposition commerciale	Rédiger	Valider prix, promesse et conditions
Analyse d'un dossier sensible	Structurer	Interpréter et décider
Action dans un outil connecté	Préparer, exécuter si autorisé	Contrôler les permissions et l'action elle-même

Le glissement à surveiller. Il ne se produit presque jamais par décision : il se produit par habitude. Une tâche commence au degré 1, donne satisfaction, et se met à passer directement en production sans que personne n'ait décidé de changer de régime. Le moment de vigilance est celui où « ça marche bien maintenant ».

À RETENIR

Préparer, recommander, décider, agir : quatre degrés, quatre exigences de contrôle. Les confondre est la cause la plus fréquente des mauvaises surprises.

GUIDE 08 · 04

Plus le système devient autonome, plus les points d'arrêt comptent

Un système agentique planifie, utilise des outils, observe le résultat et recommence. Cette autonomie est précisément ce qui le rend utile — et ce qui rend les garde-fous nécessaires.

01

limiter ce qu'il peut faire

Le périmètre des permissions se décide avant, pas pendant. C'est l'objet du Guide 07.

02

Définir où il doit demander

Lire : automatique. Créer un brouillon : automatique. Envoyer : approbation. Supprimer : bloqué.

03

Garder un moyen d'arrêt

Plus l'action est conséquente, plus la capacité de stopper, corriger ou revenir en arrière compte.

Ce que décrit Anthropic

Dans Claude, l'utilisateur peut choisir les outils activés et configurer, pour chaque action, un comportement : toujours autoriser, exiger une approbation, ou bloquer. Plutôt qu'une approbation action par action, Claude peut présenter son plan à l'avance : l'utilisateur le relit, le modifie, l'approuve — et garde la possibilité d'intervenir en cours d'exécution. ^[1]

Le principe directeur est formulé simplement : les agents doivent pouvoir travailler de façon autonome, mais les humains doivent garder le contrôle de la manière dont leurs objectifs sont poursuivis, en particulier avant les décisions à fort enjeu. ^[2]

Ce que dit — et ne dit pas — le cadre européen

L'article 14 du règlement (UE) 2024/1689 impose des exigences de contrôle humain pour les systèmes d'IA à **haut risque**. Les mesures doivent être à la mesure des risques, du niveau d'autonomie et du contexte d'utilisation du système. ^[5]

Formulation à éviter : « toute utilisation de Claude doit légalement être validée par un humain ». Formulation juste : le cadre européen illustre un principe utile — quand le risque augmente, le contrôle humain doit être proportionné et effectif ; les obligations exactes dépendent du système et du contexte juridique.

Ce qui est universel. La responsabilité ne se règle pas en choisissant un modèle ou un fournisseur. Elle se règle dans le processus de travail : qui prépare, qui vérifie, qui décide, qui peut arrêter. Ce raisonnement vaut pour tout assistant génératif.

À RETENIR

Autonomie, permissions et responsabilité se conçoivent ensemble. Ajouter de l'autonomie sans déplacer les points d'arrêt, c'est retirer du contrôle sans le dire.

GUIDE 08 · 05

Sources, périmètre et mise à jour

Sources primaires consultées le 28 août 2026.

Trois idées à emporter

1. Le niveau de contrôle dépend des conséquences d'une erreur, pas de la tâche.
2. Une validation n'est utile que si le valideur peut comprendre, contester et arrêter.
3. Autonomie, permissions et responsabilité se conçoivent ensemble.

Périmètre de ce guide

Ce guide n'est pas une analyse de la qualification d'un usage Satelliteam au regard du règlement européen sur l'IA, ni un avis juridique. Il propose un modèle de travail proportionné au risque.

À revérifier : la politique interne de validation, les processus formels applicables aux décisions concernant des personnes, et l'évolution du cadre réglementaire.

Sources primaires

[1] Trustworthy agents in practice

Anthropic Research — permissions par action, plan soumis avant exécution
anthropic.com/research/trustworthy-agents

[2] Our framework for developing safe and trustworthy agents

Anthropic — contrôle humain avant les décisions à fort enjeu
anthropic.com/news/our-framework-for-developing-safe-and-trustworthy-agents

[3] AI Risk Management Framework

NIST — cadre de gestion des risques liés à l'IA
nist.gov/itl/ai-risk-management-framework

[4] AI RMF Core

NIST AI Resource Center — fonctions Govern, Map, Measure, Manage
airc.nist.gov/airmf-resources/airmf/

[5] Règlement (UE) 2024/1689 — article 14, contrôle humain

EUR-Lex — obligations applicables aux systèmes d'IA à haut risque
eur-lex.europa.eu/eli/reg/2024/1689/oj

Point de vigilance sur les sources. Le NIST indique que l'AI RMF 1.0 est en cours de révision ; aucun calendrier de publication n'est annoncé à la date de vérification. L'article 14 du règlement européen ne s'applique qu'aux systèmes classés à haut risque : il est mobilisé ici comme modèle de bonne pratique, non comme obligation générale portant sur l'usage courant d'un assistant.