6 règles pour un agent fiable

Une checklist à coller dans le fichier d’instructions de ton agent (CLAUDE.md ou équivalent). Agenterie, octobre 2026.

Un agent IA qui écrit du code ou lance des tâches se trompe rarement de façon spectaculaire. Il se trompe en silence : une commande « réussie » qui n’a rien produit, un plan jamais relu, deux sessions qui écrasent le travail l’une de l’autre. Ces six règles viennent de l’usage quotidien d’agents en production chez Agenterie. Chacune répond à une panne réelle.

1. Prouve par la sortie réelle, pas par le code de retour

Un « exit 0 », un statut « terminé » ou une ligne « completed » en base peuvent mentir. Avant de dire qu’une tâche est faite, l’agent montre ce qu’elle a produit : les lignes écrites, le fichier daté d’aujourd’hui, le compteur qui a bougé. Une tâche longue qui écrit à la fin n’est pas en échec parce qu’elle n’a encore rien écrit à mi-parcours : on attend la fin réelle.

Test : peux-tu coller la sortie brute qui prouve le résultat ? Sinon, ce n’est pas fini.

2. Une étape qui échoue fait échouer tout le run

Interdis les filets qui transforment une panne en succès : `|| true` sur une étape critique, une valeur de repli factice, une exception avalée, un « si le dossier n’existe pas, quitter sans erreur ». Une étape d’authentification ou de données qui échoue doit sortir en erreur et déclencher une alerte.

Cas vécu : un pipeline de collecte est resté « vert » treize jours sans écrire une seule donnée. L’identifiant de connexion avait expiré, la synchronisation était suivie d’un `|| true` et un connecteur renvoyait des données factices. Un crash bruyant le premier jour aurait coûté dix minutes.

3. Un plan écrit avant toute tâche large ou irréversible

Au-delà de trois fichiers, ou dès qu’une action ne se défait pas (suppression, envoi, migration), l’agent écrit d’abord un plan : ce qu’il va faire, dans quel ordre, et surtout le **critère de fin** sous forme d’une commande ou d’un chemin qui le prouve. Il pose aussi ses hypothèses (une à trois), chacune avec la commande qui la vérifie.

Test : le plan dit-il comment on saura que c’est fini ?

4. Fais auditer le plan par un autre regard, avant la dépense

L’auteur d’un plan a des angles morts. Fais relire le plan, pas seulement le résultat, par un autre modèle ou une session sans le contexte de l’auteur. Ce relecteur est en lecture seule : s’il peut écrire, c’est un deuxième exécutant. C’est l’étape la moins chère de toute la chaîne.

Test : qui a cherché les failles du plan, et avec quelles pièces ?

5. Une session, un espace de travail

Deux agents qui modifient le même dossier se marchent dessus. Donne à chaque session qui écrit sa propre copie de travail (avec Git : un worktree par session), et ne change jamais de branche dans le dossier partagé pendant que d’autres tournent. Du code qui n’existe que sur une branche non fusionnée est du code en danger : fusionne tôt.

Test : si deux sessions tournent maintenant, touchent-elles le même dossier ?

6. L’humain garde la main sur l’irréversible

L’agent tranche seul les choix techniques et se vérifie lui-même sur les faits. Il s’arrête et demande pour quatre sujets seulement : l’argent, l’irréversible, les décisions métier, les secrets. Il ne demande jamais « je continue ? ». Les secrets ne passent jamais par la conversation ni par le dépôt de code : on indique où les lire et où les coller.

Test : cette question a-t-elle une réponse qu’une commande peut donner ? Alors l’agent la calcule.

À coller dans ton fichier d’instructions

```
## Règles de fiabilité
1. Jamais « OK » sur un signal indirect (code de retour, statut) :
   montre la sortie réelle (lignes écrites, fichier daté, compteur).
2. Interdit : "|| true" sur une étape critique, valeur de repli
   factice, exception avalée. Une étape qui échoue fait échouer
   le run et lève une alerte.
3. Tâche de plus de 3 fichiers ou irréversible : plan écrit d'abord,
   avec hypothèses vérifiables et critère de fin (une commande).
4. Le plan est relu par un autre regard, en lecture seule,
   avant d'exécuter.
5. Une session qui écrit = son propre espace de travail (worktree).
6. Question à l'humain seulement pour : argent, irréversible,
   décision métier, secret. Le reste se vérifie ou se tranche.
```

Agenterie, agenterie.fr. Texte libre de réutilisation avec mention de la source.