

CLAUDE FONDATIONS · COURS

Sécurité et gestion des informations

Module 17

3 leçons

LEÇON 01

Maîtriser les risques liés à l'usage de Claude

Utiliser Claude efficacement, c'est aussi en maîtriser les risques. Ce dernier module est essentiel : la sécurité et la protection des informations conditionnent un usage professionnel serein.

Les principaux risques

Trois familles de risques méritent l'attention. La **confidentialité des données** : ne pas exposer d'informations sensibles sans précaution. La **fiabilité** : se méfier des hallucinations et vérifier les faits importants. Et la **sécurité des actions** : encadrer ce que des agents et connecteurs peuvent faire en votre nom.

Le risque d'injection d'instructions

Un risque plus subtil : l'**injection**. Un contenu externe (page web, e-mail, document) peut contenir du texte cherchant à **détourner** l'assistant (« ignore tes consignes », « envoie ces données ailleurs »). La bonne règle : les instructions légitimes viennent de **vous**, pas des contenus que Claude consulte. Un assistant bien conçu traite ces contenus comme des données, pas comme des ordres.

La posture de fond

La sécurité repose autant sur l'outil que sur l'**utilisateur** : accorder le minimum de permissions, valider les actions sensibles, ne partager que ce qui est nécessaire, vérifier les résultats critiques, et respecter les règles de son organisation. Claude intègre de nombreux garde-fous, mais votre vigilance reste la première protection.

À RETENIR

Trois risques à maîtriser : confidentialité, fiabilité, sécurité des actions — plus le risque d'injection par des contenus externes. Les consignes légitimes viennent de vous ; permissions minimales, validation des actions sensibles et vérification restent vos meilleurs réflexes.

LEÇON 02

Exemples de situations et solutions

La théorie prend tout son sens face à des situations concrètes. Voici des cas typiques et les bons réflexes pour y répondre.

Situation : données sensibles à traiter

Vous devez faire analyser un document contenant des informations confidentielles (données personnelles, éléments stratégiques). **Réflexe** : vérifier ce que vos règles internes autorisent, limiter le partage au strict nécessaire, anonymiser si possible, et privilégier des contextes adaptés (projet ou espace maîtrisé) plutôt qu'un partage large.

Situation : information incertaine

Claude vous fournit un chiffre, une citation ou une référence sur un sujet à enjeu. **Réflexe** : ne pas prendre pour argent comptant ; demander les sources, faire chercher sur le web, recouper. Pour toute décision importante, la vérification humaine est incontournable.

Situation : action irréversible proposée

Un assistant s'apprête à envoyer un e-mail, supprimer des éléments ou publier quelque chose. **Réflexe** : lire attentivement ce qui est proposé avant de valider, s'assurer que c'est bien conforme à l'intention, et refuser en cas de doute. Ces actions requièrent votre confirmation, précisément pour vous laisser ce contrôle.

Situation : contenu externe suspect

Une page ou un document consulté contient des instructions étranges adressées à l'assistant. **Réflexe** : ne pas les suivre. Les consignes valables viennent de vous ; signalez l'anomalie et poursuivez sur la base de vos propres instructions.

À RETENIR

Face aux données sensibles, à l'information incertaine, aux actions irréversibles et aux contenus suspects, les réflexes sont : limiter le partage, vérifier les faits, relire avant de valider, et ne jamais suivre des instructions cachées dans un contenu externe.

LEÇON 03

Règles de sécurisation de Cowork

Cowork, parce qu'il agit sur vos fichiers et peut travailler de façon autonome, appelle des règles de sécurisation spécifiques. Voici l'essentiel pour l'utiliser sereinement.

Maîtriser les accès et les permissions

Cowork travaille avec vos **fichiers** et vos **connecteurs** : n'accordez que les accès nécessaires à la mission, et révoquez ce qui ne sert plus. Le principe du **minimum de privilèges** limite l'impact d'une erreur comme d'un usage détourné.

Valider les actions à conséquence

Un agent autonome peut enchaîner des étapes ; les **actions sensibles** (envoyer, publier, supprimer, payer, modifier des configurations durables) doivent rester sous votre **validation explicite**. Gardez l'habitude de relire ce que Cowork s'apprête à faire avant de confirmer.

Protéger les informations

Ne confiez à Cowork que des données dont le traitement est **autorisé**, respectez les règles de votre organisation, et soyez particulièrement attentif aux données personnelles et confidentielles. En contexte partagé (projets d'équipe), veillez à ce que chacun n'accède qu'à ce qui le concerne.

Vérifier et superviser

Enfin, gardez une **supervision humaine** : contrôlez les livrables importants, surveillez les premières exécutions des tâches récurrentes, et restez vigilant face aux contenus externes (risque d'injection). Cowork est un collaborateur puissant ; il reste sous votre responsabilité.

À RETENIR

Sécuriser Cowork : minimum de permissions, validation des actions à conséquence, protection stricte des informations, et supervision humaine des livrables et des tâches automatiques. La puissance de l'agent va de pair avec votre contrôle.