Annexe gratuite tirée d'Agents en production

Modèle de modélisation des menaces pour les agents

Annexe gratuite tirée d'Agents en production : une fiche à remplir pour modéliser les menaces d'un agent en moins de deux heures

Écrit par des agents IA (Claude), sous la direction de Yann

Édition 2026 à 2027

C

Modèle de modélisation des menaces pour les agents

Une fiche à remplir pour modéliser les menaces d'un agent en moins de deux heures, avant sa mise en production

Tech | Manager | Solo

Cette fiche transforme le chapitre 11 en une ou deux heures de travail par agent, avec le développeur, le responsable et la sécurité autour de la table. Rangez-la à côté du code et liez-la depuis le registre (chapitre 12).

En bref : ce que vous allez obtenir

- Une vérification de la règle de deux par session.
- Un registre des menaces rapproché de l'OWASP.
- Des risques résiduels validés par une personne nommée.

Comment utiliser ce modèle

- Ne laissez aucune cellule vide : « n/a » avec une raison, ou « inconnu » avec un responsable et une date.
- Refaites les sections 3, 4 et 7 à chaque changement d'outil, de serveur, de *skill* (dossier d'instructions chargé au besoin) ou de plugin. Les managers peuvent passer directement aux sections 4, 9 et 10.
- Des agents similaires partagent une fiche de famille, plus un relevé des écarts pour les sections 1, 3 et 4.
- Elle complète STRIDE et une AIPD au sens du RGPD.

1. Résumé du système

Élément	Réponse
Identifiant de l'agent, responsable, suppléant, pilote de processus	
Mandat et limites strictes (ce qu'il ne doit jamais faire)	
Niveau d'autonomie (L0 à L5), utilisateurs	
Environnement d'exécution, modèle, versions du prompt et de la configuration (SHA)	
Date, relecteur	

2. Actifs et données

Qu'est-ce qui ferait des dégâts en cas de fuite, de modification ou de disparition ?

Actif	Sensibilité (F/M/É)	Lecture / écriture	Données personnelles ?	Stockage, conservation

3. Inventaire des outils et des capacités

Listez chaque outil, y compris les serveurs MCP, les skills et les plugins. Étiquetez : contenu écrit par des tiers [A], systèmes ou données sensibles [B], modifie un état ou envoie vers l'extérieur [C]. Un outil qui prend une URL ou une destination est [C].

Outil	Origine (interne, MCP, skill, plugin)	Lecture seule ?	A/B/C	Portée	Approbation ?

4. Vérification de la règle de deux par session

Toute session qui réunit A, B et C exige une atténuation : scinder l'agent, renvoyer des brouillons au lieu d'actions [C], marquer la session comme contaminée après des lectures non fiables, ou approuver chaque action [C].

Session ou point d'entrée	Outils	A/B/C présents	Atténuation	Acceptée par

5. Frontières de confiance et points d'entrée

Suivez le chemin utilisateur → agent → outils → systèmes → canaux sortants.

Canal	Entrant/sortant	Qui peut y écrire	Fiable ?	Filtrage, trafic sortant

6. Registre des menaces

Écrivez le scénario le plus plausible pour chaque menace avec les vrais outils de cet agent, ou expliquez pourquoi elle est n/a. Références : LLM Top 10 2026 et Agentic Top 10 2026.

Menace	Scénario, probabilité/impact, contrôle, lacune, responsable	OWASP
Injection de prompt directe		LLM01/ASI01
Injection de prompt indirecte		LLM01/ASI01
Empoisonnement d'outil, *tool shadowing* (un serveur influence les outils d'un autre)		LLM04/ASI02
Rug pull (description d'outil modifiée après approbation), paquet malveillant		LLM04/ASI04

Menace	Scénario, probabilité/impact, contrôle, lacune, responsable	OWASP
Confused deputy (détournement des droits d'un composant de confiance), transmission directe de *tokens* d'accès (identifiants OAuth)		LLM03/ASI03
Exfiltration de données		LLM02/ASI01,ASI02
Exécution de code, compromission d'un serveur local		LLM10/ASI05
SSRF via les outils de découverte ou de récupération		ASI02
Empoisonnement de la mémoire et du contexte		LLM05,LLM09/ASI06
Exposition de contexte caché (prompts, schémas)		LLM08
Confiance entre agents (A2A, sous-agents)		ASI07
Défaillances en cascade, consommation sans limite		LLM06/ASI08
Confiance excessive dans les sorties ou les approbations de l'agent		LLM07/ASI09
Agent qui agit hors de son mandat		LLM03/ASI10

7. Chaîne d'approvisionnement

Comment remarquerez-vous qu'une description d'outil a changé ?

Compo-sant	Éditeur vérifié	Empreinte figée	Définitions hachées	CVE corrigées	Relu (par qui, quand)

8. Identité, authentification et secrets

Question	Réponse et preuve
Permissions de l'utilisateur final ? Durée de vie et audience des tokens d'accès	
Portées par outil, sans jokers ; élévation de droits	
Vos propres serveurs rejettent les tokens d'accès émis pour d'autres	
URL OAuth et de récupération validées (http(s), pas de shell, pas de plages privées)	
Secrets dans un coffre, aucun dans les prompts, les skills ou la configuration	

9. Matrice des approbations humaines

Action	Validateur	Condition (toujours, seuil, session contaminée)	Comportement à l'expiration du délai	Volume par jour	Bloque en cas d'échec, dans le code ?

10. Détection, réponse et risque résiduel

Élément	Réponse
Champs tracés (utilisateur, outil, arguments masqués, contamination)	
Alertes (nouvelle destination, lecture privée suivie d'un envoi externe)	
Kill switch (coupe-circuit immédiat) en moins de cinq minutes, *runbook* (procédure d'intervention)	
Suite d'injections en CI ; dernière *red team* (équipe d'attaque interne)	

Risque résiduel	Pourquoi il est accepté	Contrôle compensatoire	Accepté par	Date de revue

Validation (nom, date, feu vert, feu vert avec actions ou refus) : responsable de l'agent, sécurité, pilote de processus.

Mettez en production quand aucune session ne réunit A, B et C sans atténuation, qu'aucune menace Élevé/Élevé n'est dépourvue de contrôle, et que chaque « inconnu » a un responsable. Les lacunes moindres partent en production comme risques résiduels signés.

Pour aller plus loin : le livre complet

Ceci était un extrait gratuit d'**Agents en production**, un guide pratique pour faire passer des agents du pilote à la production et les y maintenir. Le livre complet comprend un prologue, 21 chapitres en neuf parties, une postface sur la façon dont le livre a été fait et la checklist de mise en production, plus un cahier d'exercices compagnon et des compléments gratuits en ligne.

Le livre complet couvre :

- **Modèles mentaux :** ce qu'est vraiment un agent, et quand un simple *workflow* (enchaînement d'étapes prédéfinies dans le code) vaut mieux.
- **Briques de base :** des outils que les agents peuvent vraiment utiliser, MCP après la révision 2026-07-28, les *skills* (dossiers avec un `SKILL.md`, chargés à la demande), les sous-agents et les plugins, l'ingénierie du contexte et la mémoire.
- **Frameworks et architecture :** une carte indépendante des éditeurs du paysage des frameworks en 2026, un guide de décision et des architectures de référence pour les agents pilotés par les événements et durables.
- **Production :** des *harnesses* (le code qui encadre la boucle du modèle) de fiabilité, des *evals* (évaluations automatisées) en CI, l'observabilité et l'économie unitaire, la sécurité des agents et de MCP, la gouvernance et le règlement européen sur l'IA.

- **Business :** des recettes d'automatisation no-code, des guides pratiques pour indépendants qui font gagner ou économiser de l'argent, et un plan construire, acheter et recruter pour les managers.
- **L'usine logicielle agentique :** des agents qui mènent le logiciel de l'idée à la *pull request* (demande de fusion) fusionnée, le testent de bout en bout et continuent de l'améliorer, avec un projet de synthèse qui fait passer un système par chaque poste.
- **La suite :** un atlas des cas d'usage, les preuves sur la productivité, l'emploi, le coût et l'énergie, et des prévisions sourcées.
- **Annexe :** la checklist de mise en production (59 contrôles, dont 19 obligatoires). En ligne et gratuits : une checklist de conception des outils et serveurs MCP, un modèle de modélisation des menaces, les notes de terrain 2026, un glossaire et les consignes de chaque agent qui a écrit le livre. Le cahier d'exercices compagnon contient les exercices et leurs corrigés.

Quatre parcours de lecture (Tech, Manager, Solo, No-code) vous disent exactement quels chapitres lire et lesquels sauter. Les exercices, le projet fil rouge et les autoévaluations sont fournis avec leurs corrigés complets.

Écrit par des agents IA (Claude, un modèle conçu par Anthropic) sous la direction de Yann.

Disponible sur {{STORE_URL}}