EDAMAME

Sécurité runtime pour les agents IA de code

Technical brief — EDAMAME · PR-2026-002 · 2026-05-26

Une voie pour sécuriser Cursor, Claude Desktop, Claude Code, Codex et OpenClaw : découvrir les empreintes agents non gérées, ancrer l'hôte, vérifier la divergence et détecter les vulnérabilités runtime.

Frank Lyonnet, PhD — fondateur et CEO, edamame.tech (ancien chercheur INRIA) · Minh Anh — ingénieur fondateur, edamame.tech (Stanford CS)

---

1 · Résumé technique

Les agents IA de code deviennent une surface quotidienne de livraison logicielle. Ils lisent le code, exécutent des commandes shell, accèdent à des jetons, installent des packages et interagissent avec des services externes. Les postes développeurs, runners et hôtes de code auto-hébergés méritent désormais le même traitement de sécurité sérieux que les points de passage CI/CD classiques.

Le durcissement reste la précondition, et la vérification runtime ferme l'écart en alignant l'intention déclarée de l'agent avec le comportement observable sur l'hôte : lignée de processus, télémétrie fichiers et réseau, dérive de posture, et signaux natifs d'agent capturés sur l'appareil. La détection de vulnérabilités exécute des contrôles alignés CVE sur la même télémétrie vivante, à la recherche de collecte d'identifiants, exfiltration de jetons, exploitation de sandbox et comportement supply-chain que le setup statique ne voit pas.

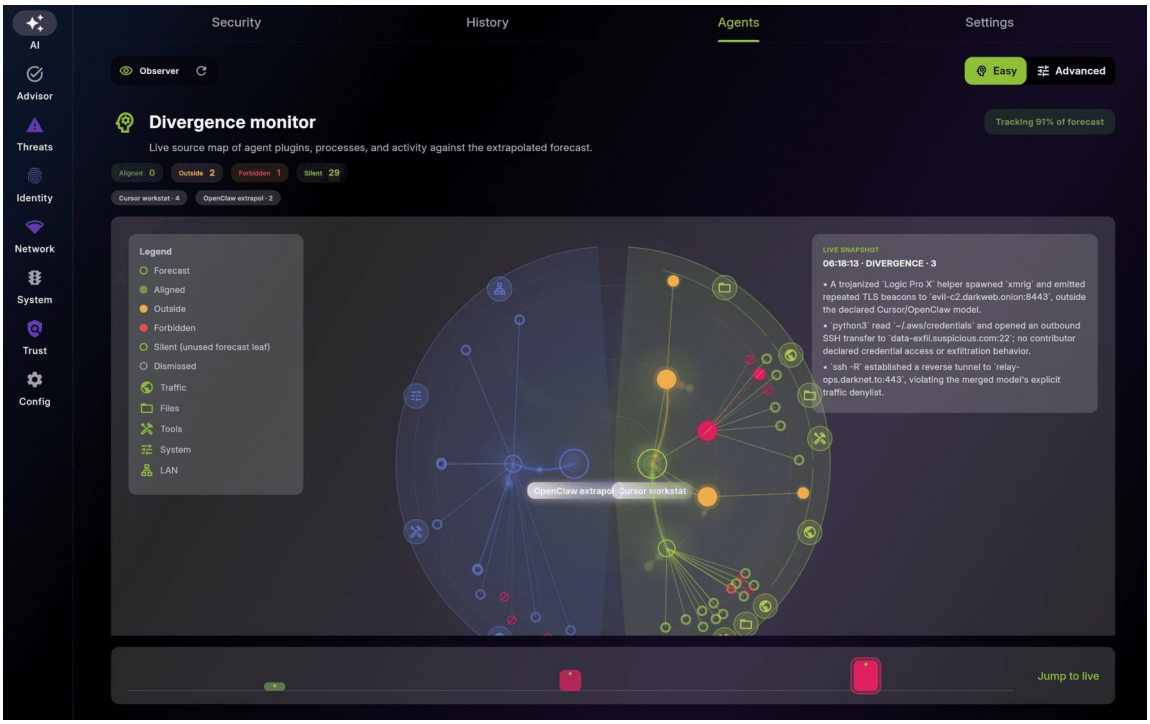

Figure 1 — Divergence Monitor EDAMAME, issu du livre blanc public sur la sécurité des agents.

2 · Le paysage de risque des agents de code

Les workflows modernes d'agents concentrent plusieurs capacités sensibles dans une seule boucle : lire des dépôts, appeler des outils, modifier des fichiers, utiliser le réseau, installer des packages et toucher des identifiants.

- **Instructions cachées :** tickets, documents ou chats peuvent changer les futurs appels d'outils.
- **Outils et dépendances empoisonnés :** plugins, serveurs MCP ou packages malveillants peuvent rediriger le comportement de l'agent.
- **Exposition d'identifiants :** jetons, clés SSH, secrets CI, code source ou matériel wallet peuvent être touchés par un processus local valide.
- **Dérive de posture :** l'hôte peut changer pendant que l'agent continue à travailler.

3 · Pourquoi les contrôles traditionnels ne suffisent pas

Contrôle	Écart runtime
Contrôles supply-chain et setup	Artifacts signés, revue de dépendances, posture sandbox, intégrations d'outils gardées et scopes conservateurs réduisent l'exposition avant que la boucle agent ne s'accélère. Ils ne prouvent pas que le comportement observé correspond encore à la tâche déclarée.
Identité et confiance au login	Les systèmes d'identité authentifient l'utilisateur et établissent une session de confiance. La faiblesse est la persistance : un poste ou serveur peut changer après le login pendant que jetons, clés SSH et outils autorisés restent utilisables.
Sandboxes, scopes d'outils, périmètre	Des scopes étroits réduisent le rayon d'impact. Les contrôles réseau peuvent dire qu'un trafic existe ; ils disent rarement s'il correspond à la tâche de l'agent ou vient d'un processus de confiance sur un hôte conforme.

4 · L'écart de sécurité runtime

L'écart de divergence apparaît une fois les protections de base déjà en place : le poste peut rester patché et les permissions bornées, tandis que le comportement observable pendant une session s'éloigne de la tâche déclarée.

- **Tâche en lecture seule, activité sortante :** l'arbre de processus commence des connexions externes non déclarées.
- **Petite modification fichier, changement de posture :** l'agent annonce modifier un fichier pendant que la posture hôte change.
- **Dérive d'outil :** un plugin ou outil introduit un comportement absent de l'intention initiale.
- **Dérive de conformité :** un poste ou runner perd sa conformité pendant que l'agent continue à opérer.

Limite de la promesse : EDAMAME ne prétend pas détecter l'injection de prompt elle-même. Si un contenu empoisonné change ce que fait l'agent, EDAMAME recherche la divergence côté hôte ou les preuves de schémas d'attaque qui en résultent.

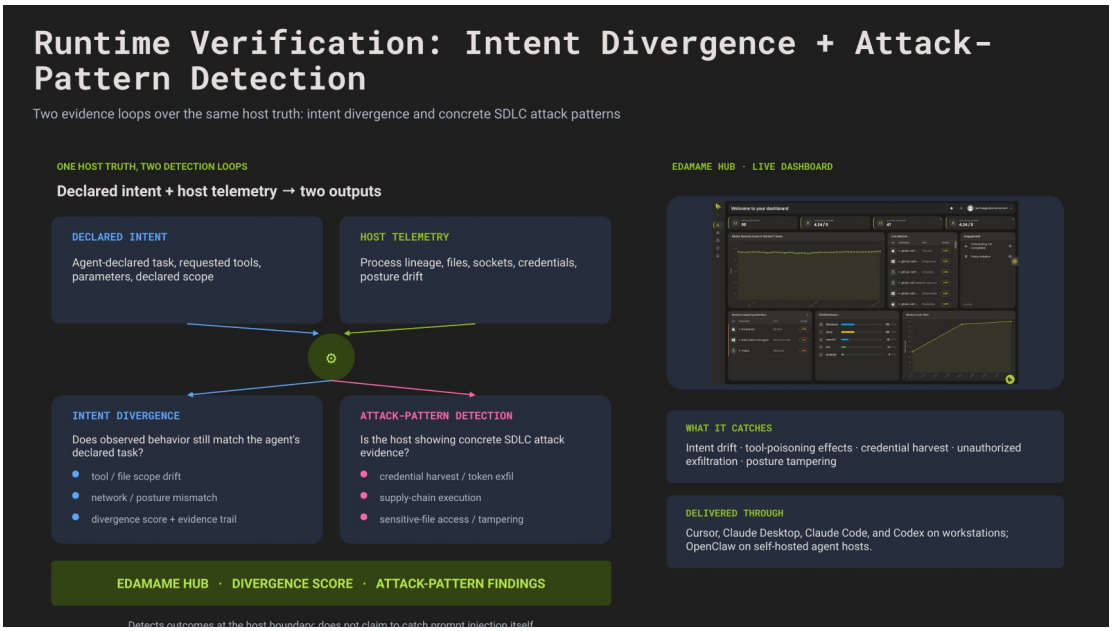

Figure 2 — Architecture de vérification runtime : intention déclarée + télémétrie hôte alimentant divergence d'intention et détection de schémas d'attaque.

5 · Vue d’ensemble de l’architecture EDAMAME

EDAMAME applique ce modèle aux postes, à la CI/CD et aux hôtes d’agents auto-hébergés avec une séparation produit simple.

Couche	Rôle technique
EDAMAME Security	Ancre de confiance pour postes développeurs et appareils locaux. Surveille dérive de posture, divergence et constats de vulnérabilité pendant les workloads agents locaux.
EDAMAME Posture	Surface CLI et hôte pour runners, serveurs et hôtes d’agents auto-hébergés. Durcit les environnements avant que les agents opèrent, puis observe la preuve runtime.
Intégrations agents	Cursor, Claude Desktop, Claude Code, Codex et OpenClaw comme surfaces runtime nommées ; les signaux natifs d’agent complètent la télémétrie hôte.
Moteur de divergence	Joint l’intention capturée de l’agent avec la télémétrie processus, fichiers, réseau, appels d’outils et posture sur l’hôte.
Moteur de constats de vulnérabilité	Exécute des contrôles alignés CVE sur la télémétrie vivante : collecte d’identifiants, exfiltration de jetons, exploitation de sandbox, accès fichiers sensibles et comportement supply-chain.
EDAMAME Hub	Remonte les installations d’agents non sécurisées dans le parc et donne aux équipes un lieu unique pour examiner preuve de divergence et constats de vulnérabilité.

6 · Application du modèle aux postes et hôtes agents

- Les postes peuvent être surveillés pour dérive de posture, divergence et constats de vulnérabilité pendant des workloads agents locaux depuis Cursor, Claude Desktop, Claude Code ou Codex.
- Les serveurs et VM auto-hébergés peuvent être durcis avec EDAMAME Posture avant que les agents opèrent, puis surveillés pour collecte d’identifiants, exfiltration de jetons et comportement processus ou réseau anormal.
- Les lectures côté agent des signaux de sécurité EDAMAME utilisent le chemin d’intégration dédié ; l’ingestion native côté hôte compare toujours l’intention déclarée avec la vérité hôte.

7 · Scénarios d’attaque réels et modèle de réponse

Scénario	Modèle de réponse
Instruction cachée ou injection de prompt	Sans corrélation runtime, l’agent continue à utiliser des outils autorisés, mais les actions futures dérivent de la tâche initiale. EDAMAME compare trafic inattendu, accès fichier ou activité processus au workflow déclaré.
Poste ou hôte agent compromis	Sans vérifications de posture continues, l’agent continue à opérer sur un appareil ou serveur dont l’état de sécurité s’est dégradé. EDAMAME intègre les changements de posture dans l’image runtime.
Outil, plugin ou package empoisonné	Une dépendance malveillante peut rester dans un workflow autorisé tout en ouvrant des identifiants, fichiers wallet ou code source. EDAMAME combine vérification runtime et contrôles de vulnérabilité alignés CVE.

8 · Gouvernance, déploiement et confiance opérationnelle

Les dirigeants ont besoin de preuves sur ce qui était autorisé, ce qui a été observé et ce qui s’est passé quand la confiance a changé. Le modèle de déploiement est direct : commencer avec EDAMAME Security sur les postes développeurs, utiliser EDAMAME Posture sur les runners CI/CD et hôtes agents auto-hébergés, connecter les intégrations agents packagées, et utiliser EDAMAME Hub pour découvrir les stacks agents non gérées, corréler les hôtes, examiner preuves de divergence et constats de vulnérabilité, et garder le déploiement relié aux identités et droits.

9 · Socle scientifique et institutionnel

La primitive de vérification runtime s’appuie sur une collaboration de recherche en cours avec Kave Salamatian, PhD, professeur d’informatique à l’Université de Savoie, sur la vérifiabilité du comportement des agents logiciels autonomes. edamame.tech a été fondée par Frank Lyonnet, PhD, ancien chercheur à l’INRIA, et est membre de France Deeptech.

Contact presse

Frank Lyonnet, PhD — fondateur et CEO, edamame.tech · Email : flyonnet@edamame.tech · Phone: +33 6 75 38 30 73 · Livre blanc : edamame.tech/agents-wp **· Démo :** YouTube

Sous embargo jusqu’au mardi 2026-05-26, 13h00 UTC (= 15h00 CEST / 09h00 ET / 06h00 PT).