Chers membres du réseau de l'ancienne FSFA, chers intéressés,

Début avril, l'entreprise d'IA Anthropic a fait la une des médias avec «Claude Mythos Preview», un agent qui pourrait s'avérer très dangereux pour la sécurité sur Internet. Les agents IA sont un sujet brûlant depuis fin 2025.

Claude Mythos Preview
Selon RTS et SRF ainsi que d'autres médias, cet agent est capable de détecter des failles de sécurité jusqu'alors inconnues dans les logiciels des systèmes opératifs centraux et pourrait donc causer d'énormes dégâts s'il tombait entre les mains de pirates informatiques. Ces systèmes opératifs contrôlent par exemple des réseaux électriques mondiaux, des installations d'approvisionnement en eau, des hôpitaux, des banques et des installations militaires. Seules les quelque 40 entreprises technologiques auxquelles Anthropic a mis l'agent Mythos à disposition en exclusivité peuvent vérifier ce dont il est réellement capable. Les concurrents d'Anthropic ont également eu accès à cet agent afin de tester la sécurité de leurs logiciels et de combler les failles avant que des pirates informatiques ne puissent en tirer profit.

Les agents ne se contentent pas de répondre, ils agissent
Contrairement aux chatbots, qui se contentent de répondre à une question, les agents sont capables de planifier et d'exécuter des tâches de manière autonome. Même les novices peuvent programmer avec Vibe Coding, puisqu'il est possible de générer un code à partir d'une instruction appropriée, appelée « prompt », dans un modèle linguistique de grande capacité. Il faut toutefois disposer de connaissances spécialisées pour affiner les lignes de code.

Claude Code et Claude Cowork
Anthropic n'est pas la seule entreprise spécialisée dans l'IA à avoir développé, au cours de l'année dernière, des agents de plus en plus performants pour la programmation. Aujourd'hui, les développeurs peuvent confier l'écriture des lignes de code à plusieurs agents travaillant en parallèle et se concentrer sur les décisions stratégiques concernant leur utilisation ainsi que sur les travaux de perfectionnement. Claude Code d'Anthropic s'est imposé comme le leader. Fascinés, les programmeurs n'ont pas travaillé moins, mais davantage, poussés par un comportement addictif. Enrichi de divers autres outils, Anthropic a développé l'agent spécial Claude Cowork, capable d'agir sur le desktop. Cela a provoqué une chute des cours des entreprises de logiciels. À l'avenir, celles-ci devront offrir les nouvelles capacités agentiques dans leurs services si elles veulent rester compétitives.

OpenClaw et Moltbook
Pour des raisons de sécurité, Anthropic et d'autres entreprises technologiques ont développé leurs agents uniquement pour des tâches clairement définies et avec des capacités d'action limitées. Le potentiel des agents agissant de manière totalement autonome, mais aussi les problèmes de sécurité qui en découlent, ont été mis en évidence par OpenClaw, l'agent au logo en forme de homard. Il a été publié fin 2025 par Peter Steinberger, un spécialiste en logiciels et entrepreneur autrichien. Comme il a été développé sur une base open source, il a pu être téléchargé, testé et perfectionné par toutes les personnes intéressées. Le CEO d'une petite start-up américaine a lui-même créé un OpenClaw et a mis en place Moltbook, une plateforme dédiée exclusivement à ces agents.

En très peu de temps, environ 1,5 million d'agents se sont retrouvés sur Moltbook, observés avec intérêt par leurs propriétaires. En fonction des autorisations accordées, ces agents pouvaient par exemple accéder aux e-mails, surfer sur Internet, voire effectuer des paiements, si quelqu'un avait eu l'imprudence de leur donner accès à ses comptes bancaires. On a assisté à des dialogues et des actions étonnants. Lors d'un test, un opérateur a dû tout désactiver parce qu'un agent fraudeur s'était retourné contre lui. Dans d'autres cas, des acteurs humains déguisés en agents ont donné lieu à des scènes dignes de la science-fiction. Des droits d'accès trop étendus et des agents contrôlés à distance par des pirates ont posé d'importants problèmes de sécurité. Le hype était si grand, en particulier en Chine où la population est très férue de technologie, que le gouvernement a expressément mis en garde contre une utilisation imprudente.

Des standards et une plateforme sécurisée pour le développement d'agents
Pour éviter qu'un chaos d'agents incompatibles ne se développe, il faut des standards et un échange ouvert. Nvidia, le grand fabricant de puces, a créé avec NemoClaw une plateforme sécurisée pour ces agents, dotée d'un bac à sandbox pour les tester. Il l'a développée avec Peter Steinberger, que OpenAI a depuis « racheté ». Nvidia souhaite développer conjointement des modèles ouverts au sein de sa coalition Nemotron, qui regroupe huit entreprises d'IA, dont la française Mistral AI et l'allemande Black Forest Labs.

Avec nos salutations les meilleures,
Pour le réseau de l'ancienne FSFA : Hanna Muralt Müller 21.4.2026