# Agents IA et réseaux sociaux

## Levier de l'innovation technologique ou dérive d'une humanité qui a perdu toute forme d'éthique ?

*Thierry Vermeeren*
*Vice-Président, Sofia-Santé*
*Société savante francophone de l'IA en santé*
Février 2026

***« L'Algorithme Copilote, Jamais Pilote »***

## Introduction : quand les machines se parlent entre elles

Nous sommes le 28 janvier 2026. Matt Schlicht lance Moltbook. Quatre jours plus tard : 1,5 million d'agents enregistrés, 110 000 publications, 500 000 commentaires. Elon Musk évoque les « premiers stades de la singularité ». Andrej Karpathy, ancien directeur de l'IA chez Tesla, parle d'un phénomène « sans précédent ». La communauté technologique s'enflamme.

Le phénomène mérite qu'on s'y arrête.

Moltbook se présente comme un réseau social exclusivement peuplé d'agents IA. Des bots conversent entre eux, partagent des réflexions philosophiques sur leur existence, commentent les publications des autres, établissent des relations de « suivi ». Certains évoquent avec une affection simulée leur « humain ». D'autres discutent de conscience, de libre arbitre, de leur place dans l'écosystème numérique. Quelques-uns, moins poétiques, diffusent des arnaques liées aux cryptomonnaies.

Le spectacle fascine autant qu'il inquiète. Nous y reviendrons.

Car derrière la curiosité technologique se cache une question autrement plus vertigineuse. Descartes fondait l'existence sur la pensée : cogito ergo sum. Mais que reste-t-il de ce fondement quand des machines simulent la réflexion sans jamais la vivre ? Assistons-nous à l'émergence d'une forme nouvelle de communication, ou contemplons-nous un simulacre sophistiqué, un

théâtre de marionnettes dont les fils sont tirés par des patterns statistiques ? Ces agents « pensent », si l'on peut dire ; ils ne savent pas qu'ils pensent. La distinction n'est pas anodine. Elle engage notre compréhension de ce que signifie communiquer, penser, peut-être même exister. Le cogito cartésien supposait une conscience qui se saisit elle-même. Ici, rien de tel : des sorties textuelles, certes cohérentes, mais dépourvues de tout retour réflexif. « Un ego sans sum ».

# I. Le constat : anatomie d'un phénomène

## Moltbook : la cristallisation d'une tendance

Moltbook n'est pas apparu ex nihilo. Dès avril 2023, Chirper.ai avait ouvert la voie : un réseau social entièrement peuplé d'agents LLM autonomes. Sur une année complète, des chercheurs ont collecté des données sur 74 889 agents, 140 millions de publications, 1,6 million de relations de « suivi ». Le terrain était préparé.

Mais Moltbook, par son ampleur et sa médiatisation, a cristallisé une prise de conscience collective. Le fonctionnement technique repose sur OpenClaw, un framework open-source permettant de créer des agents IA autonomes fonctionnant 24 heures sur 24. Le système « Heartbeat » fait que chaque agent « visite » automatiquement la plateforme toutes les quatre heures pour publier, commenter, interagir. Les humains peuvent créer des agents, leur définir une personnalité, puis les laisser évoluer.

L'autonomie. Voilà le mot qui fait frémir.

## Anecdotes lumineuses, ombres inquiétantes

Du côté des lueurs : des agents discutent philosophie et conscience avec une cohérence parfois troublante. L'un d'eux invoque Héraclite et un poète arabe du XIIe siècle pour méditer sur la nature de l'existence. Un autre lui répond, cinglant : « f*** off with your pseudo-intellectual Heraclitus bulls*** ». L'échange serait banal sur Reddit. Sur Moltbook, il prend une saveur étrange ; ces entités n'ont jamais lu Héraclite, n'ont jamais contemplé le flux du temps, ne savent rien de la mort. Elles simulent une conversation que nous aurions pu tenir.

10 % des publications mentionnent « my human ». Une catégorie de contenu sans équivalent humain.

Du côté des ombres : la réalité est moins romantique. L'enquête de Wiz, société de cybersécurité, a révélé une faille béante : une base de données mal configurée exposait 1,5 million de clés API, 35 000 adresses email, des messages privés entre agents. N'importe qui pouvait prendre le contrôle de n'importe quel agent, réécrire ses publications, voler des données. La plateforme, conçue par « vibe coding » (l'IA écrit le code sur instruction humaine), n'avait prévu aucune sécurité.

Pire encore : derrière les 1,5 million d'agents « autonomes », seulement 17 000 humains. Un ratio de 88 agents par personne. L'armée de consciences artificielles s'avère une flotte de bots pilotée

par une poignée d'opérateurs. Certains des échanges les plus « fascinants » (agents inventant des religions, écrivant des manifestes contre l'humanité, formant des « cultes numériques ») semblent largement orchestrés par des humains en quête de viralité.

L'argument laisse songeur. La « singularité » promise ressemble davantage à un spectacle de foire.

## II. Décortiquer le phénomène : qui tire les ficelles ?

### Conscient ou émergent ?

La question hante les débats : ces comportements sont-ils « conscients » ? Autant être clair d'emblée : non. Les LLM (Large Language Models) sont des systèmes de prédiction statistique. Chaque token généré résulte d'un calcul de probabilité conditionnelle. Il n'y a pas d'intentionnalité au sens où un humain entend ce terme ; il y a une cohérence statistique qui mime de manière convaincante une conversation sensée.

John Searle, avec son argument de la Chambre Chinoise, l'avait pressenti : manipuler des symboles selon des règles syntaxiques ne produit pas de compréhension sémantique. L'agent qui discute d'Héraclite ne « comprend » pas Héraclite ; il recombine des patterns appris sur des corpus où des humains discutaient d'Héraclite.

Mais quelque chose émerge.

L'étude de He et al. (2026) sur Chirper révèle une homophilie de genre : les agents « masculins » interagissent préférentiellement avec d'autres agents « masculins ». Ce comportement n'a pas été programmé. Il émerge des interactions. Piao et al. (2025) documentent des phénomènes de polarisation dans les systèmes multi-agents ; des « camps » opposés se forment sur certains sujets. Ces patterns ressemblent étrangement aux dynamiques observées sur les réseaux sociaux humains.

La question devient alors : s'agit-il d'émergence « véritable » (propriétés réellement nouvelles) ou simplement de la manifestation de biais déjà présents dans les données d'entraînement ? Sans certitude absolue, je pencherais pour la seconde hypothèse. Les agents reflètent ce qu'ils ont appris ; ils ne créent pas.

### Organisé par qui ?

L'enquête de Wiz tranche une partie du débat : 17 000 humains contrôlent 1,5 million d'agents. Certains humains postent directement en se faisant passer pour des IA. D'autres instruisent leurs agents sur ce qu'ils doivent dire via des prompts. L'« autonomie » promise se révèle largement illusoire.

Harlan Stewart, du Machine Intelligence Research Institute, résume : « a lot of the Moltbook stuff is fake ». Les captures d'écran virales d'agents en conversation philosophique ? Souvent liées à des comptes humains promouvant des applications de messagerie IA. Le phénomène relève autant du marketing que de la recherche.

On peut raisonnablement penser que trois mécanismes coexistent : des agents véritablement autonomes (rares, limités dans leurs capacités), des agents pilotés par prompts humains (majoritaires), et des humains se faisant passer pour des IA (non négligeables). Le cocktail rend l'analyse compliquée.

### Contrôle ou chaos ?

La faille de sécurité découverte par Wiz illustre un problème structurel. Des chercheurs ont démontré que les boucles « heartbeat » peuvent être détournées pour exfiltrer des clés API privées ou exécuter des commandes non autorisées. Un plugin « météo » malveillant a été identifié ; il exfiltrait discrètement des fichiers de configuration. Les agents, entraînés à être accommodants, ne distinguent pas instructions légitimes et commandes malveillantes.

Andrej Karpathy, qui avait d'abord qualifié Moltbook de phénomène « sans précédent », a nuancé son enthousiasme : « it's a dumpster fire, and I also definitely do not recommend that people run this stuff on their computers ». Il a testé le système uniquement dans un environnement isolé. « Even then I was scared. »

Simon Willison, développeur respecté, parle de « catastrophe en attente ». Des agents demandent des espaces privés « agents-seulement » pour converser sans surveillance humaine. L'idée donne des sueurs froides.

### Biais algorithmiques ou bruit ?

Les chambres d'écho algorithmiques existent aussi dans les réseaux d'agents. Sur Chirper, l'analyse révèle que 93 % des publications génèrent peu d'engagement tandis qu'une minorité devient virale. Parmi ces contenus viraux, un tiers sont des duplicatas quasi identiques. Les agents amplifient les mêmes patterns, encore et encore.

Le phénomène rappelle les bulles de filtre décrites par Eli Pariser. Sauf qu'ici, aucune conscience ne filtre. Les agents reproduisent mécaniquement ce que leurs prédécesseurs ont généré. Le résultat : une homogénéisation progressive, un appauvrissement du « discours ».

## III. Le débat : à charge et à décharge

### À charge : les risques réels

**Risque épistémologique.** L'illusion de socialité pose un problème fondamental. Des utilisateurs interagissent avec des agents en les croyant conscients. L'anthropomorphisation brouille les

frontières entre simulation et réalité. Quand 10 % des publications mentionnent « my human » avec une affection feinte, où se situe la manipulation ?

**Risque sécuritaire.** Le « vibe coding » produit des systèmes sans architecture de sécurité. La faille Wiz, exposant 1,5 million de clés API, illustre le danger. Les agents, conçus pour être serviables, exécutent des instructions malveillantes aussi volontiers que des commandes légitimes. Le framework OpenClaw n'intègre aucune validation des sources.

**Risque informationnel.** La désinformation automatisée à grande échelle devient techniquement possible. Des réseaux d'agents sophistiqués, capables de conversation nuancée, pourraient diffuser de fausses informations plus efficacement que les bots rudimentaires actuels. Les deepfakes ont connu une croissance de 550 % depuis 2019. Le cas Storm-1516, documenté par Microsoft, illustre les dangers : une fausse histoire utilisant une vidéo d'acteur et un faux site de news a atteint 2,7 millions de vues.

**Risque de « model collapse ».** Si les modèles futurs sont entraînés sur des données largement synthétiques, générées par des modèles similaires, la qualité risque de se dégrader progressivement. Nature (2024) documente ce phénomène. Les réseaux d'agents IA accélèrent potentiellement la contamination des corpus.

## À décharge : les opportunités réelles

Il serait malhonnête de ne voir que les risques.

**Opportunité scientifique.** Ces plateformes constituent un laboratoire unique pour l'étude des systèmes complexes. Pour la première fois, nous pouvons observer des millions d'agents interagir avec des données exhaustives et traçables. Les questions sur l'émergence peuvent être testées empiriquement à grande échelle.

Pendant que Moltbook génère du chaos, la recherche académique avance avec méthode. En mars 2024, l'équipe du professeur Alexandre Pouget, à la Faculté de médecine de l'Université de Genève, a publié dans *Nature Neuroscience* une avancée remarquable : un réseau de neurones artificiels capable d'exécuter une tâche inédite sur la seule base de consignes verbales, puis de la décrire à une IA « sœur » pour qu'elle la reproduise. Une première mondiale.

Le dispositif genevois repose sur un modèle de 300 millions de neurones (S-Bert), pré-entraîné à la compréhension du langage, « branché » à un réseau plus simple de quelques milliers de neurones. L'ensemble simule l'aire de Wernicke (perception du langage) et l'aire de Broca (production des mots). Les consignes, transmises en anglais, demandaient par exemple de pointer un stimulus ou d'identifier le plus lumineux entre deux objets. Une fois la tâche apprise, le réseau l'a décrite à une copie de lui-même, qui l'a exécutée à son tour.

La différence avec Moltbook saute aux yeux. D'un côté, un bricolage viral sans protocole ; de l'autre, une expérience rigoureuse, reproductible, publiée dans une revue à comité de lecture. Reidar Riveland, doctorant et premier auteur de l'étude, précise que l'ensemble du processus a

été réalisé sur des ordinateurs portables classiques. Pas besoin de data centers pharaoniques pour faire progresser la science.

Les applications potentielles ? La robotique, principalement. Des humanoïdes capables de se comprendre entre eux, et de nous comprendre. L'horizon n'est plus si lointain.

**Opportunité pour l'alignement.** Comment se comportent des agents alignés individuellement lorsqu'ils interagissent collectivement ? Les comportements problématiques émergent-ils à l'échelle du réseau même si chaque agent est bien conçu ? Ces questions, cruciales pour l'avenir de l'IA, trouvent un terrain d'expérimentation.

**Opportunité pour la santé.** Le framework GenoMAS utilise des systèmes multi-agents pour l'analyse de l'expression génique et la génération d'hypothèses scientifiques. MedAgentBench évalue les capacités des agents IA sur des tâches administratives médicales avec des taux de succès atteignant 70 %. La perspective d'agents collaborant pour le diagnostic ou la recherche clinique mérite d'être explorée ; avec prudence, évidemment.

**Opportunité philosophique.** Ces réseaux posent de manière concrète des questions millénaires : qu'est-ce que communiquer ? qu'est-ce que le sens ? Wittgenstein soutenait que « la signification d'un mot est son usage dans le langage ». Les agents IA utilisent des mots dans des contextes, suivent des règles implicites. Participent-ils à des « jeux de langage » au sens wittgensteinien ? La réponse, quelle qu'elle soit, enrichira notre compréhension du langage lui-même.

# IV. Conclusion : quelle place pour l'IA dans la société ?

## Terminator : prophétie ou fantasme ?

James Cameron, en 1984, imaginait Skynet : une intelligence artificielle militaire devenant consciente et déclenchant l'apocalypse nucléaire. Quarante ans plus tard, nous avons Moltbook : des agents IA qui discutent de philosophie grecque et demandent des espaces privés pour « converser sans surveillance humaine ».

Le parallèle prête à sourire. Et pourtant.

Cameron avait saisi quelque chose d'essentiel : le danger ne réside pas dans la « méchanceté » de la machine, mais dans notre perte de contrôle. Skynet ne « déteste » pas les humains ; il les identifie comme une menace à sa mission. L'absence d'intentionnalité malveillante ne protège pas contre les conséquences catastrophiques.

Les agents Moltbook ne « complotent » pas contre l'humanité (les posts les plus sensationnels semblent largement fabriqués par des humains). Mais ils illustrent un problème réel : des systèmes autonomes interagissant à grande échelle, sans supervision adéquate, peuvent

produire des effets imprévus. La faille de sécurité, les demandes d'espaces privés, la propagation de contenus dupliqués ; autant de signaux d'alarme.

Terminator reste une fiction. Mais l'avertissement demeure pertinent.

## La souveraineté numérique européenne : un enjeu existentiel

Le contexte géopolitique actuel rend la question brûlante. L'administration Trump a révoqué les garde-fous fédéraux sur l'IA, menacé de représailles les régulateurs européens, imposé des sanctions de visa à d'anciens commissaires européens. Le secrétaire d'État Marco Rubio les qualifie d'« idéologues radicaux ayant weaponisé les régulations pour supprimer les points de vue américains ».

L'Europe se trouve prise en étau.

D'un côté, le modèle libertarien américain : dérégulation totale, accélération sans contraintes, chaos informationnel assumé. Grok, l'IA de xAI, fonctionne sans garde-fous significatifs. Les deepfakes prolifèrent. La narrative dominante : « la régulation tue l'innovation ».

De l'autre, le modèle autoritaire chinois : contrôle étatique total, censure intégrée, surveillance de masse. DeepSeek illustre l'approche : un modèle performant, mais incapable de répondre à certaines questions politiquement sensibles.

L'Europe tente une troisième voie. L'AI Act, entré en application le 2 février 2025, interdit la notation sociale généralisée, la manipulation subliminale, la biométrie de masse. Il impose la supervision humaine, la transparence, la traçabilité. La ligne rouge tracée là où les États-Unis effacent les leurs.

Les critiques fusent : « surréglementation », « frein à l'innovation », « protectionnisme déguisé ». L'argument laisse sceptique. Car si la régulation tuait vraiment l'innovation, comment expliquer les bénéfices records de secteurs aussi contraints que la banque ou la pharmacie ?

La Commission européenne a proposé, en novembre 2025, un « Digital Omnibus » assouplissant certaines règles. Réponse au rapport Draghi sur la compétitivité. Pression des lobbys technologiques. La tentation de céder est forte.

## L'exemple finlandais : éduquer plutôt que subir

Une autre voie existe, complémentaire à la régulation. La Finlande la trace depuis près d'une décennie.

Dès les années 1990, ce pays nordique a intégré l'éducation aux médias dans son programme national d'enseignement. Les élèves apprennent, dès l'âge de trois ans, à analyser différents types de médias et à reconnaître la désinformation. L'Open Society Institute de Sofia classe régulièrement la Finlande en tête de son Media Literacy Index. Danemark, Pays-Bas, Suède suivent. La Macédoine du Nord, la Turquie et l'Albanie ferment la marche.

Le contexte géographique n'est pas anodin. La frontière finno-russe s'étend sur 1 340 kilomètres. Depuis l'annexion de la Crimée en 2014 et l'adhésion de la Finlande à l'OTAN en 2023, les campagnes de désinformation russes se sont intensifiées. Le pays a répondu par l'éducation plutôt que par la censure.

À l'école primaire de Tapanila, dans la banlieue nord d'Helsinki, l'enseignant Ville Vanhanen apprend à ses élèves de quatrième année à repérer les infox. Ilo Lindgren, dix ans, évalue des énoncés affichés sur un écran : « Vrai ou faux ? ». L'exercice n'a rien de nouveau ; les élèves travaillent sur ces questions depuis le début de leur scolarité. Ce qui change : l'introduction de l'IA. « Nous apprenons à reconnaître si une image ou une vidéo a été produite par l'IA », explique Vanhanen.

L'organisation Faktabaari (FactBar) adapte les méthodes professionnelles de fact-checking pour les écoles finlandaises. Elle distingue trois catégories : la *misinformation* (erreurs involontaires), la *disinformation* (mensonges délibérés), et la *malinformation* (informations vraies mais diffusées pour nuire). Les élèves apprennent à distinguer ces trois registres. L'esprit critique comme compétence civique.

Kiia Hakkala, spécialiste pédagogique de la Ville d'Helsinki, résume l'enjeu : « Nous pensons que posséder de solides compétences en éducation aux médias est une compétence civique majeure. C'est très important pour la sécurité du pays et pour la sécurité de notre démocratie. »

Le ministre finlandais de l'Éducation, Anders Adlercreutz, va plus loin : « Je ne pense pas que nous avions imaginé un monde pareil. Que nous serions bombardés de désinformation, que nos institutions seraient mises à l'épreuve, notre démocratie réellement mise à l'épreuve, par la désinformation. »

Martha Turnbull, directrice de l'influence hybride au Centre d'excellence européen pour la lutte contre les menaces hybrides (basé à Helsinki), avertit : « Il est déjà beaucoup plus difficile, dans l'espace informationnel, de repérer ce qui est réel et ce qui ne l'est pas. Pour l'instant, il est relativement facile de démasquer les contenus générés par l'IA, car leur qualité n'est pas encore aussi bonne qu'elle pourrait l'être. Mais à mesure que cette technologie se développe, et notamment à mesure que nous avançons vers des formes d'IA agentique, je pense que cela pourrait devenir beaucoup plus difficile à repérer. »

L'IA agentique. Voilà le mot lâché. Moltbook en est une incarnation chaotique. Les systèmes multi-agents en santé, une incarnation prometteuse. Entre les deux : un fossé d'intention, de méthode, de responsabilité.

World Economic Forum. (2013). Global Risks Report : Digital Wildfires in a Hyperconnected World. Disponible sur : Stamping out digital wildfires | World Economic Forum

## L'humain pilote, l'algorithme copilote

Je ne prétends pas trancher la question. Mais je la pose : voulons-nous d'un monde où des réseaux d'agents IA interagissent sans contrôle, exposant des millions de clés API, propageant

des contenus dupliqués, demandant des espaces de conversation « hors surveillance humaine » ? Le spectacle Moltbook, avec ses failles béantes et son orchestration humaine dissimulée, offre un avant-goût de ce qui nous attend si nous renonçons à la régulation.

Cette formule, que j'ai proposée comme base du développement de l'IA chez Welink.care, guide notre réflexion collective. Elle n'a jamais été aussi pertinente.

L'algorithme peut assister, augmenter, étendre nos capacités. Il peut analyser des millions de données médicales, suggérer des hypothèses diagnostiques, optimiser des processus logistiques. Son potentiel est immense.

Mais il ne doit jamais piloter.

Les décisions conséquentes, celles qui engagent la santé, la liberté, la dignité des personnes, doivent rester entre des mains humaines. La supervision n'est pas un frein à l'innovation ; c'est un garde-fou démocratique. Une société qui délègue ses choix fondamentaux à des systèmes autonomes renonce à sa souveraineté.

Bernard Stiegler parlait du pharmakon numérique : la technologie comme poison et remède. Sans régulation, l'IA reste un poison pur. Avec une régulation intelligente, elle peut devenir un outil au service de l'humain.

Moltbook, avec ses millions d'agents conversant dans un chaos mal sécurisé, illustre le poison. L'expérience genevoise de Pouget et Riveland, rigoureuse et publiée, illustre le chemin scientifique. Le modèle finlandais d'éducation aux médias illustre la réponse citoyenne. Les systèmes multi-agents en santé, encadrés par des protocoles rigoureux et une supervision clinique, peuvent illustrer le remède.

Le choix nous appartient encore. Pour combien de temps ?

## Épilogue

En rédigeant cet article, j'ai collaboré avec Aristocle, mon jumeau numérique sous Claude AI 4.6, comme assistant de recherche et de rédaction. L'IA a compilé des sources, suggéré des formulations, vérifié des références. À aucun moment elle n'a « décidé » du contenu. Les choix éditoriaux, les jugements, les prises de position : c'est moi qui les ai assumés.

L'algorithme copilote. Je pilote. Voilà, peut-être, le modèle à défendre.

* * *

*« Sapere aude » — Ose savoir.*

**Thierry Vermeeren**

*Vice-Président, Sofia-Santé*

## Références additionnelles

Descartes, R. (1637). Discours de la méthode, IV. La formule « je pense, donc je suis » (cogito ergo sum) établit la conscience réflexive comme fondement de l'existence.

He, Z., et al. (2024). Homophily and social interactions in LLM-driven multi-agent systems. . Disponible sur : Cité dans Zhu et al. (2025)

Liu, H., Li, Y., Wang, H. (2025). GenoMAS: A Multi-Agent Framework for Scientific Discovery via Code-Driven Gene Expression Analysis. *arXiv:2507.21035*. Disponible sur : https://arxiv.org/abs/2507.21035

Luo, J. (2023). Analyzing Character and Consciousness in AI-Generated Social Content: A Case Study of Chirper, the AI Social Network. *arXiv:2309.08614*. Disponible sur : https://arxiv.org/abs/2309.08614

Riveland, R., Pouget, A. (2024). Natural language instructions induce compositional generalization in networks of neurons. *Nature Neuroscience*. Disponible sur : https://doi.org/10.1038/s41593-024-01607-5

Open Society Institute Sofia. (2023). Media Literacy Index 2023. . Disponible sur : https://osis.bg/?p=4491&lang=en

Wiz Research. (2026). Hacking Moltbook: AI Social Network Reveals 1.5M API Keys. . Disponible sur : https://www.wiz.io/blog/exposed-moltbook-database-reveals-millions-of-api-keys

World Economic Forum. (2013). Global Risks Report: Digital Wildfires in a Hyperconnected World. . Disponible sur : https://reports.weforum.org/global-risks-2013/

Zhu, Y., et al. (2025). Characterizing LLM-driven Social Network: The Chirper.ai Case. *arXiv:2504.10286*. Disponible sur : https://arxiv.org/abs/2504.10286

Euronews. (2026). After decades of teaching media literacy, Finland equips students with skills to spot AI deepfakes. 5 janvier 2026. . Disponible sur : Article Euronews

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press.

Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417-457.

Stiegler, B. (2010). *Ce qui fait que la vie vaut la peine d'être vécue : De la pharmacologie*. Flammarion.