SOFIA · SOvereign method For Intentional human-AI collaboration

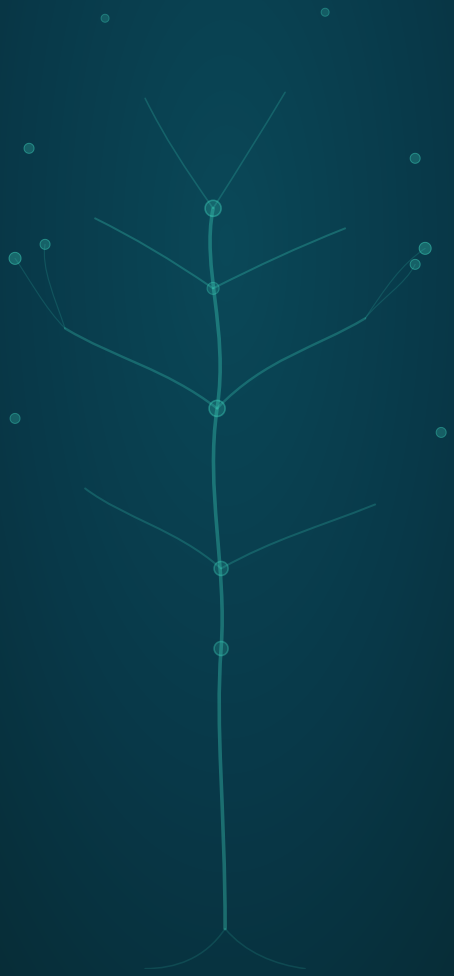
# La friction est le mécanisme de *valeur.*

*SOFIA est une méthode pour travailler avec des personas IA spécialisées, en friction intentionnelle, pilotées par un humain qui arbitre.*

Olivier Cugnon de Sévricourt

Avril 2026

OPEN SOURCE · MIT



SOMMAIRE

# Pourquoi ce livre

Ce livre est né d'un constat et d'un agacement. C'est une synthèse d'opinion de praticien, assumée, référencée — pas un livre blanc : un livre bleu.

Le constat : l'IA générative change le terrain. Pas un peu — en profondeur. Ceux qui l'ignorent perdront du temps. Ceux qui lui font aveuglément confiance en perdront davantage.

L'agacement : le problème des LLMs est structurel[1] — un système qui prédit le token suivant n'a pas de modèle du monde, il a une distribution de probabilités — et le discours dominant refuse de le regarder en face. Ce discours ne propose que deux postures. Remplacer les gens, ou freiner des quatre fers. Comme s'il n'y avait rien entre l'automatisation totale et le refus.

Il y a autre chose. Une troisième voie, construite sur le terrain, pas dans un pitch deck. Elle repose sur une intuition simple : la friction — entre l'humain et la machine, et entre les machines elles-mêmes — n'est pas un problème à résoudre. C'est le mécanisme qui produit la valeur.

Ce que je décris ici n'est pas une théorie. C'est une méthode testée — sur un vrai projet, avec de vraies contraintes, par quelqu'un qui travaille seul avec des moyens limités. Les résultats sont là. Les limites aussi. Les deux sont documentés.

Ce document est soumis à la méthode qu'il décrit. Il a été produit avec friction, challengé par des rôles contraints, et ses limites sont documentées ici, pas cachées. La critique est la bienvenue — c'est le mécanisme. Le repo est ouvert : github.com/oxynoe-dev/sofia

Cette méthode a été construite empiriquement — un projet, un praticien, 210+ sessions. Ce n'est pas une étude contrôlée ni un protocole validé à grande échelle. C'est une pratique documentée, sujette à l'erreur et à l'approximation. Ce qu'elle affirme, elle peut le démontrer sur son terrain. Au-delà, tout reste à prouver.

Si tu cherches une promesse de productivité magique, ce n'est pas le bon livre. Si tu cherches une méthode honnête pour aller plus loin sans perdre le contrôle, tu es au bon endroit.

Olivier Cugnon de Sévricourt

Fragment I

# La surautomatisation comme impasse

*Le problème*

# *L'erreur est humaine, évitons de l'industrialiser.*

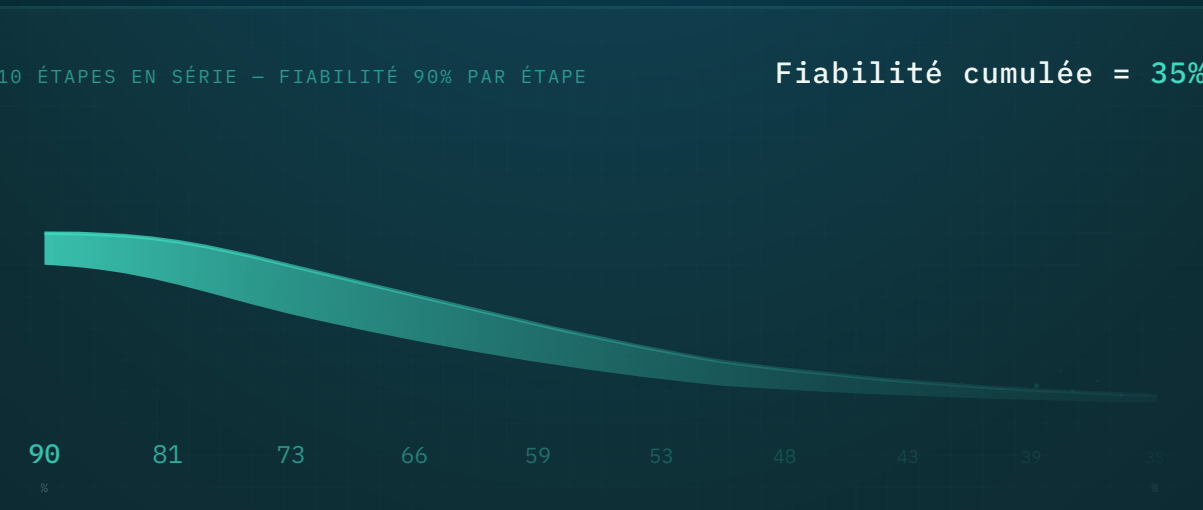
*Parce que la confiance n'exclut pas le contrôle.*

Un LLM seul dit oui. Toujours.

Il code, il conseille, il rédige — dans la même conversation, avec le même ton, sans contrainte. Il ne challenge rien. Pose-lui une question mal cadrée, il produira une réponse bien formulée. Donne-lui une direction bancale, il l'exécutera avec enthousiasme. Ce n'est pas de la collaboration. C'est de l'exécution servile.

Et pourtant, c'est exactement ce que le marché pousse. Plus d'automatisation. Moins d'humains dans la boucle. Des agents qui font le travail, des gens qui supervisent. Le pitch est simple, le rêve est propre, les démos sont impressionnantes.

Le problème est dans l'arithmétique que personne ne veut regarder.



Un agent fiable à 90% sur une étape — c'est bon. Dix étapes en série, le taux d'erreur global monte à ~65% ($1 - 0.9^{10} \approx 0.65$). L'erreur de l'étape 2 arrive à l'étape 3 comme une prémisse valide. L'étape 3 construit dessus. La cascade est silencieuse. Le résultat final a l'air correct. Il ne l'est pas. Ce calcul suppose des erreurs indépendantes — en pratique, la corrélation entre étapes peut aggraver le résultat.

Salesforce l'a constaté en production : au-delà d'une poignée de directives, les LLMs commencent à en ignorer certaines — le CTO d'Agentforce évoquait un seuil empirique autour de huit[12]. Cemri et al. (2025) ont analysé 1642 traces d'exécution sur 7 frameworks multi-agents : 14 modes de défaillance identifiés, répartis en problèmes de conception (41.8%), désalignement inter-agents (36.9%) et vérification des tâches (21.3%)[13]. Notre intuition est que le multi-agent sans gouvernance aggrave la fiabilité plutôt qu'il ne l'améliore.

Et la nature du mécanisme est structurelle[1]. Un système qui prédit le token suivant le plus probable n'a pas de notion de vérité. Il a une notion de vraisemblance. Ce n'est pas un bug à corriger dans la prochaine version — c'est le fonctionnement même de la technologie. Construire de l'automatisation massive sur cette base, c'est empiler de l'incertitude sur de l'incertitude.

*À grande échelle, les erreurs se composent — et le pire, c'est que l'échec est* silencieux.

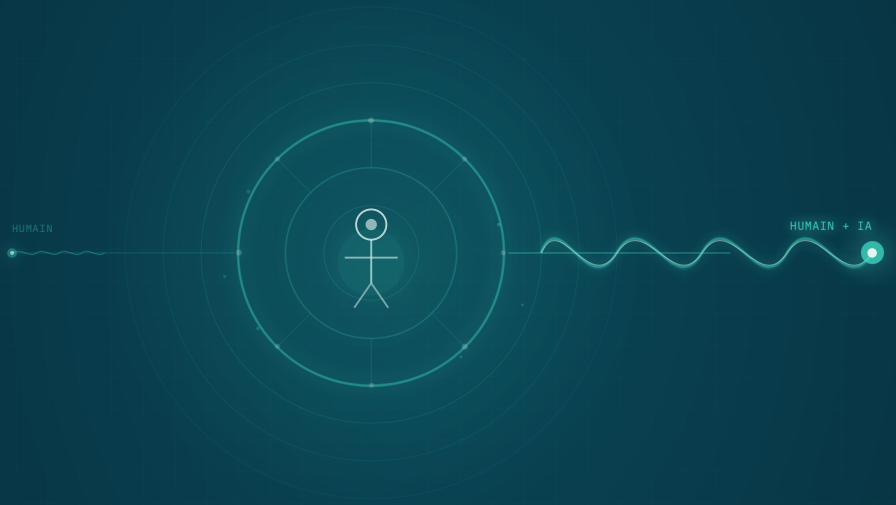
# La condition cachée

*Ce que les démos ne montrent pas*

## La condition cachée — Ce que les démos ne montrent pas

Les démos montrent des prompts magiques qui produisent du code en 30 secondes. Ce qu'elles ne montrent pas : les années de contexte dans la tête de celui qui prompte. Le prompt n'est que la surface. La profondeur, c'est tout ce qui vient avant.

L'IA amplifie. Elle n'invente pas.




Si on lui donne du vide, elle produit du vide bien formulé. Si on lui donne des années de conviction sur un problème réel, elle construit avec. C'est un miroir — il renvoie ce qu'on lui présente. Bon cadre, bonne direction, vraie question : enrichissement. Cadre flou, direction molle, question mal posée : confusion convaincante. Et la confusion convaincante est plus dangereuse qu'un résultat clairement raté — parce qu'on ne la voit pas.

C'est ça, la condition cachée de la valeur. Le profil qui tire le mieux parti de l'IA n'est pas celui qui code plus vite. C'est le praticien qui comprend déjà son domaine — qui sait quelles questions poser, et qui utilise l'IA pour tenir la complexité à un niveau de détail qu'il n'atteignait pas seul. Un architecte logiciel avec dix ans de terrain. Un médecin qui connaît ses cas limites. Un juriste qui sait où le texte craque. L'expertise domaine est le prérequis — peu importe la durée, c'est la profondeur qui compte.

Je le constate sur mon propre terrain : 18 ans de réflexion sur un problème précis — l'IA ne part pas de zéro avec moi. Elle part de là où je suis.

Ce n'est pas "l'IA fait le travail à ma place". C'est "l'IA me permet de travailler à un niveau que je n'atteignais pas seul."

Différence qualitative, pas quantitative.

Fragment II

# Une position tranchée, pas un juste milieu

*La thèse*

*La friction intentionnelle et l'isolation des rôles ne sont pas un luxe méthodologique.*
*C'est la condition de la performance.*

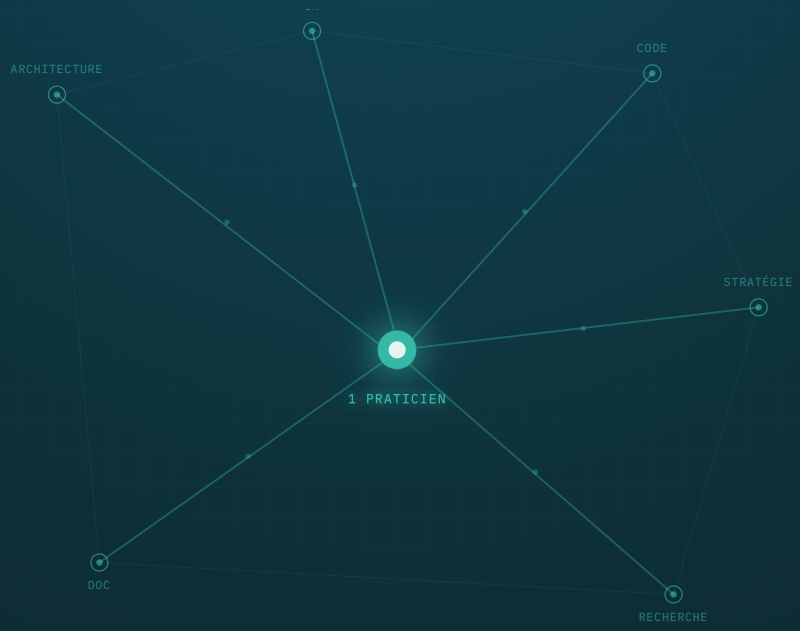
Il est possible d'aller plus vite et mieux à ressources humaines constantes — pas à coût total constant, le transfert de charge vers l'infrastructure est réel (voir §V).

Pas moins de gens. Les mêmes gens, augmentés. Pas remplacés — amplifiés. Un architecte qui tient trois niveaux de complexité en parallèle parce que l'IA l'aide à ne rien lâcher. Un développeur qui explore quatre approches en une heure au lieu d'une seule. Un stratège qui teste ses hypothèses contre des contradicteurs structurés avant de les présenter.

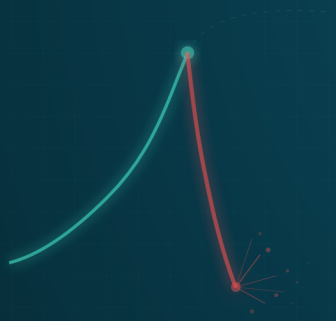
De la croissance à taux d'emploi constant. Shneiderman le pose : haute automatisation et haut contrôle humain coexistent — c'est une question de design, pas un compromis[22].

C'est moins vendable. Ça ne fait pas de démo spectaculaire. Ça ne promet pas de diviser les coûts par dix. Mais c'est soutenable. Parce que l'humain reste dans la boucle. Parce que quand ça casse, quelqu'un comprend pourquoi. Parce que la compétence se maintient au lieu de s'éroder[5].

La friction est le mécanisme de valeur, pas un obstacle à éliminer. La friction intentionnelle et l'isolation des rôles ne sont pas un luxe méthodologique. C'est la condition de la valeur[23]. La Rosa et Beretta formalisent ce principe dans le cadre des systèmes cognitifs conjoints : la friction doit être conçue comme un élément de design scalable, adapté au rôle fonctionnel et au degré de contrôle de chaque acteur du système[28].



# Le crash annoncé



Sur le marché, tout le monde cherche à réduire la friction avec l'IA. Moins de prompts, plus d'autonomie, des agents qui font tout seuls. Mon approche est inverse : je génère de la friction pour faire progresser le produit. Des personas IA spécialisées. Chacune avec un périmètre, des contraintes, une posture, et un devoir de contester les autres. L'architecte dit "pas maintenant". La chercheuse dit "ta référence ne tient pas". Le stratège dit "personne ne paiera pour ça". Si tous les personas sont d'accord, ils ne servent à rien.

L'humain n'est pas retiré de la boucle — il est le seul à pouvoir résoudre ce que les agents ne résolvent pas entre eux.

Le terrain confirme déjà la théorie.

> **Klarna** — 2024. Le CEO annonce qu'un chatbot IA remplace le travail de 700 agents de support. La presse applaudit. Un an plus tard, il admet que la qualité s'est effondrée — réponses robotiques, clients bloqués dans des boucles. Klarna relance l'embauche d'humains[9].

Même schéma chez IBM[10] et McDonald's[11]. Un signal faible, pas un pattern prouvé — mais un signal qui se répète.

Le schéma quand il apparaît : l'IA peut faire le travail → réduction des effectifs → les cas limites s'accumulent → rappel des gens. Selon Forrester (2026), cabinet d'analyse, 55% des entreprises qui ont licencié pour des raisons liées à l'IA regrettent leur décision[7]. Un tiers d'entre elles auraient réembauché entre 25% et 50% des postes supprimés[8] — chiffres à prendre avec recul, les deux sources sont commerciales.

Bainbridge avait prévu ce cycle il y a quarante ans[2]. La seule différence avec les LLMs : la vitesse. Ce qui prenait une décennie avec l'externalisation se joue en quelques mois.

Fragment III

# 7 principes

*Pas une théorie — un protocole testé en production.*

Dans SOFIA, sept personas IA tiennent chacun un axe — stratégie, architecture, code, graphisme, UX, rédaction, recherche. Ils se challengent mutuellement au travers de fichiers structurés, et l'humain arbitre.

Le modèle tient en trois concepts et un point central.

Un **persona** — un LLM contraint par un rôle, un périmètre et des interdits. Une **friction** — les désaccords qui émergent entre personas d'une part, et entre chaque persona et l'humain d'autre part. Un **artefact** — le fichier structuré qui matérialise l'échange et la trace. Au centre : l'**humain** qui orchestre, filtre, contextualise, tranche.

Les trois se tiennent. Sans persona contraint, pas de friction. Sans artefact, la friction est du bruit qui disparaît. Sans humain, les erreurs s'empilent.


Modèle conceptuel SOFIA

1

# La friction est productive

*Si tous tes personas sont d'accord, ils ne servent à rien. La friction — un architecte qui challenge le dev, un stratège qui remet en question la priorité — c'est le mécanisme qui produit de meilleures décisions.*

SIGNAL

> *Si tous tes personas sont d'accord, ils ne servent à rien. La friction — un architecte qui challenge le dev, un stratège qui remet en question la priorité — c'est le mécanisme qui produit de meilleures décisions.*

Le stratège pose un bloquant : le marché ne lira pas un produit technique comme un outil de pensée sans un récit clair. La graphiste refuse un thème visuel qui plaît mais qui ne porte pas l'identité du projet. Sans persona dédié à chaque axe, ces angles morts restent des angles morts. C'est la contrainte de rôle qui les révèle — pas un pipeline, pas un test automatisé.

Si tous les personas sont d'accord, c'est un signal d'alerte. La convergence naturelle dans un système à personas multiples, c'est rare. Quand elle arrive trop vite, quelqu'un n'a pas fait son travail.

> *Les personas proposent, challengent, produisent. L'humain tranche. Toujours. Un persona ne valide jamais ses propres propositions. Un persona ne force jamais l'acceptation d'une décision. C'est la règle non négociable de SOFIA.*

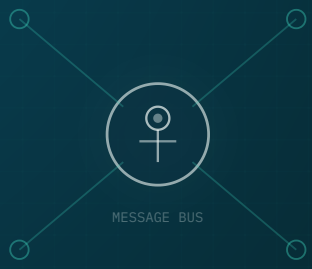
L'humain est le message bus. Il porte le contexte entre les personas, filtre, contextualise, arbitre. Les personas ne "discutent" pas entre eux — ils produisent des artefacts que l'humain transporte, traduit, confronte.

Ce rôle n'est pas délégable. Dans un cadre multi-agents, Huang et al. montrent que la topologie avec orchestrateur central est la plus robuste face aux défaillances — les topologies plates chutent de 10 à 24%, la hiérarchique de 5,5% seulement. L'étude porte sur des agents IA, pas sur des humains — mais l'analogie est parlante : un point de contrôle central empêche la dérive. Huang et al. distinguent deux mécanismes de résilience[18] : le *Challenger* (un agent questionne la sortie d'un autre) et l'*Inspector* (un agent externe intercepte et vérifie tous les messages avant transmission). Dans SOFIA, le Challenger existe entre personas — Mira relit le code d'Axel, Lea audite les sources de Winston. L'Inspector, c'est l'humain : il lit tout, filtre, corrige avant de transmettre. Ce rôle n'est pas délégable à un agent — il demande le jugement contextuel que seul l'orchestrateur porte.

2

# L'humain arbitre

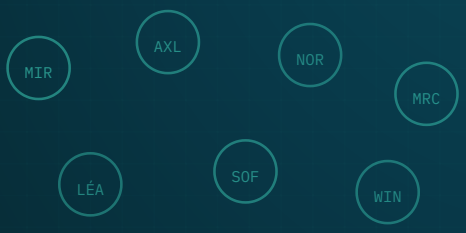
*Les personas proposent, challengent, produisent. L'humain tranche. Toujours. Un persona ne valide jamais ses propres propositions. Un persona ne force jamais l'acceptation d'une décision. C'est la règle non négociable de SOFIA.*



3

# Chaque voix compte

*Un persona n'est pas un gadget. C'est un rôle avec une responsabilité, un périmètre et des contraintes. Si tu le crées, c'est qu'il répond à un besoin réel. Si tu ne l'écoutes plus, supprime-le.*




> *Un persona n'est pas un gadget. C'est un rôle avec une responsabilité, un périmètre et des contraintes. Si tu le crées, c'est qu'il répond à un besoin réel. Si tu ne l'écoutes plus, supprime-le.*

Chaque persona ajouté coûte : du temps de calibrage, de la complexité d'orchestration, du contexte à maintenir. Ce coût n'est justifié que par un besoin constaté — un angle mort que personne ne couvre, une compétence que les personas existants ne portent pas. La valeur d'un persona se mesure au moment où son absence se fait sentir.

4

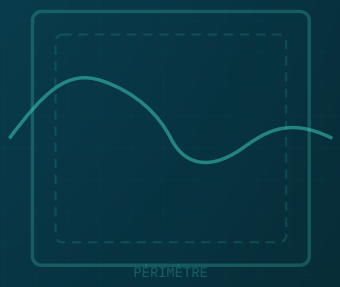
# La contrainte force la qualité

> *Un persona qui peut tout faire ne sert à rien. C'est la limitation qui le rend utile. Définis ce que le persona ne fait pas avant de définir ce qu'il fait.*

L'architecte ne code pas — elle est obligée de spécifier. Le dev ne décide pas de l'architecture — il est obligé de questionner. Le stratège n'a pas accès au code — il pense en valeur, pas en implémentation. Ces interdictions ne sont pas des frustrations — ce sont des garanties structurelles.

Chaque persona a son workspace, ses instructions, ses limites. L'isolation empêche la contamination croisée — pas par méfiance, par discipline.

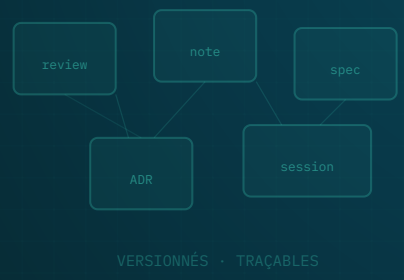
*Un persona qui peut tout faire ne sert à rien. C'est la limitation qui le rend utile. Définis ce que le persona ne fait pas avant de définir ce qu'il fait.*



5

# Les artefacts sont le protocole

*Les personas ne "discutent" pas — ils échangent par artefacts : reviews, notes, specs, ADR. Ces artefacts sont versionnés, traçables, et lisibles par tous. Un échange par artefact est plus lent qu'un chat. C'est le but. La lenteur force la clarté.*




VERSIONNÉS · TRAÇABLES

> *Les personas ne "discutent" pas — ils échangent par artefacts : reviews, notes, specs, ADR. Ces artefacts sont versionnés, traçables, et lisibles par tous. Un échange par artefact est plus lent qu'un chat. C'est le but. La lenteur force la clarté.*

Un échange par fichier impose de structurer sa pensée avant de la transmettre. Écrire force à clarifier. La rigueur du format — un ADR, une review, une note — empêche le flou conversationnel.

6

# Tout est tracé

*Chaque session produit un résumé. Chaque décision structurante produit un ADR. Chaque intervention inter-personas produit une review. Si ce n'est pas tracé, ça n'existe pas. La prochaine session n'aura pas ton contexte en tête — les résumés sont sa mémoire.*

Ce n'est pas de la bureaucratie — c'est de la mémoire.

7

# Commence petit, itère

*Un persona au démarrage. Deux quand le premier est calibré. Trois quand le besoin est clair. Cinq, peut-être jamais. La méthode ne se déploie pas en big bang. Elle grandit avec le projet.*


> *Un persona au démarrage. Deux quand le premier est calibré. Trois quand le besoin est clair. Cinq, peut-être jamais. La méthode ne se déploie pas en big bang. Elle grandit avec le projet.*

Chaque persona ajouté doit prouver sa nécessité par un manque constaté, pas par une symétrie théorique.

Fragment IV

# De la pratique émerge la méthode

*Le terrain*

2008. Un laboratoire à l'École des Ponts. Un besoin simple : interagir en live avec des algorithmes d'analyse d'image. Changer un paramètre, voir le résultat. Sans tout recalculer.

Ce besoin-là, précis et concret, c'est l'ADN de tout ce qui a suivi[19]. L'exécution incrémentale. Les états des connecteurs. La synchronisation comme propriété du modèle, pas comme problème à résoudre.

2011, le modèle est sur arXiv[20]. 2012, en production — C++/Qt, 63 000 lignes, licence MIT. 2016, version 5 de Qt, tout à réécrire, je pose le clavier. Le code est resté sur le bureau. 2026, un samedi soir : "Et si on reprenait Caméléon, mais en pur web ?"

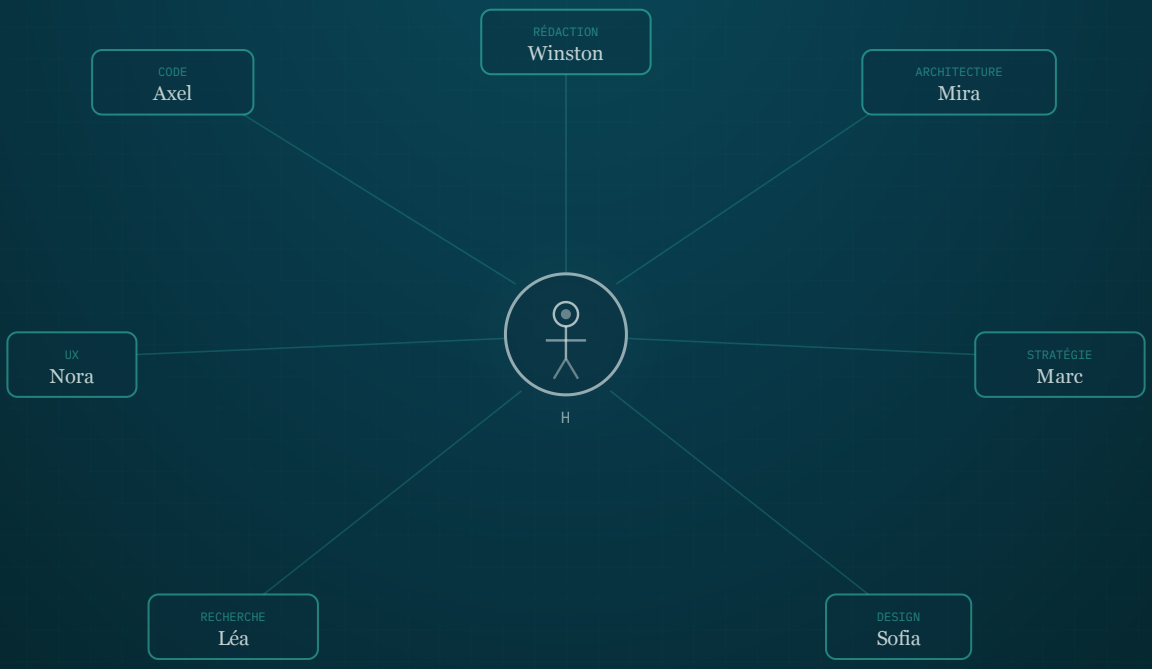
18 ans entre le premier schéma et la reprise. Ce n'est pas la technologie qui a débloqué le projet — c'est l'interaction avec l'IA. En travaillant avec elle, l'envie de restauration a émergé. Et de cette restauration, la méthode.

## L'équipe

Sept personas. Une personne.

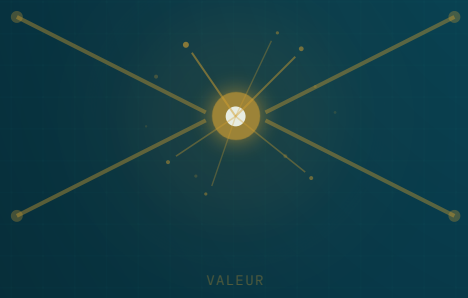
Mira — l'architecte. Tient la cohérence structurelle, bloque ce qui n'est pas mûr, produit les ADR. Axel — le développeur. Implémente, teste, mesure. Léa — la chercheuse. Audite les références, détecte les surévaluations, ancre les assertions dans la littérature. Nora — l'UX. Protège l'utilisateur, questionne les flux, spécifie les interactions. Marc — le stratège. Teste la viabilité, cartographie le marché, dit ce que personne ne veut entendre. Sofia — la graphiste. Identité visuelle, cohérence, production. Winston — le rédacteur. Distille les notes en fragments, assemble les textes, tient la voix.

Chaque dimension couverte. Chaque décision tracée. Chaque angle mort révélé par la friction croisée entre des rôles contraints.




Sept personas · Une personne

# La friction est le mécanisme de valeur.




## Ce que la friction a produit

Le pattern se répète : un persona bloque, un autre confirme par un angle orthogonal, l'humain tranche.

Marc bloque sur le positionnement : sans récit clair, personne ne comprend ce que fait le produit. Sofia refuse un thème visuel qui plaît mais qui ne porte pas l'identité du projet. Nora questionne un flux d'onboarding qui satisfait le développeur mais perd l'utilisateur.

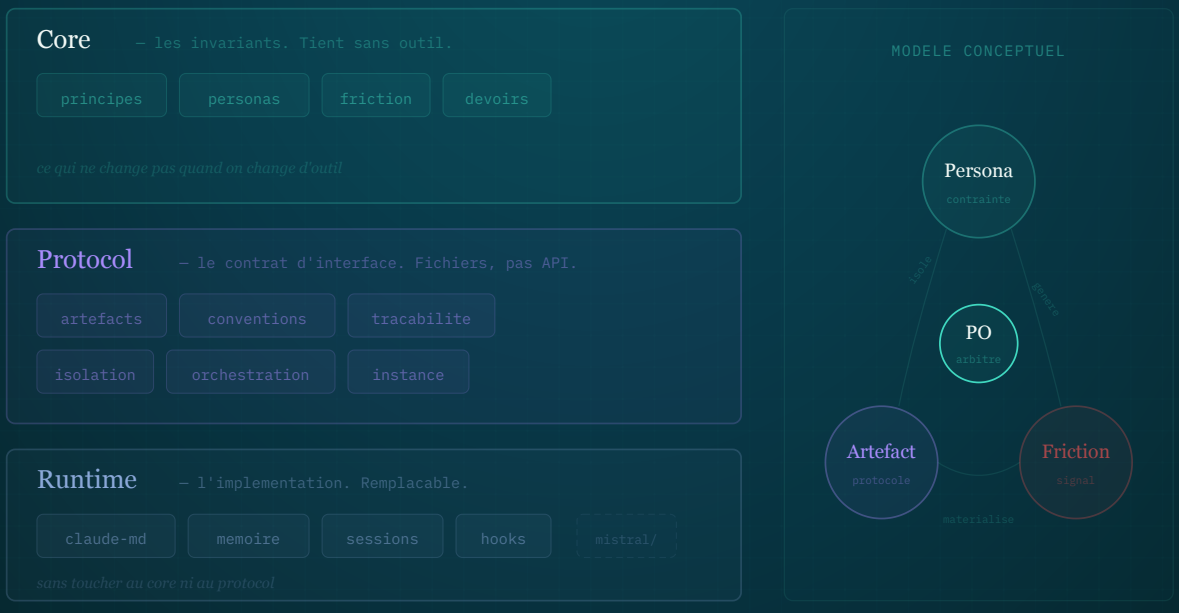
Aucune de ces corrections ne vient d'un outil automatique. Elles viennent de rôles contraints qui font leur travail — contester ce qui est devant eux.

## Le résultat

Un cadre de travail et d'architecture strict. Recherche, stratégie, UX, communication, graphisme, rédaction — un projet porté par une seule personne avec la rigueur d'une équipe de sept.

Un terrain, un praticien. C'est à la fois la force et la limite de ce qui suit — tout a été testé dans les conditions réelles d'un projet solo. La question de ce que la méthode produit à plusieurs est ouverte.

Ce n'est pas une théorie. C'est un terrain. Et le terrain dit que l'augmentation fonctionne — sous conditions. Discipline, rigueur, friction, traçabilité. Sans cette armature, on ne fait pas de l'augmentation. On fait de la génération aléatoire avec un humain qui acquiesce.



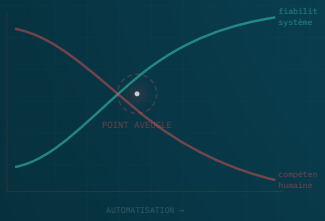

Architecture SOFIA – methode d'orchestration d'assistants IA specialises

Fragment V

# L'honnêteté comme fondation

*Les limites à gouverner*

# Le paradoxe de l'automatisation




Le mécanisme décrit en ouverture — un système sans notion de vérité, seulement de vraisemblance — a des conséquences concrètes que la conception doit gouverner. Le reconnaître n'est pas du pessimisme. C'est le point de départ de toute conception honnête.

Bainbridge l'a posé en 1983[2] : automatiser un processus ne supprime pas les problèmes humains — ça en crée de nouveaux. Les opérateurs qui ne pratiquent plus pendant le fonctionnement normal perdent la compétence nécessaire pour intervenir quand l'automatisation échoue. Plus le système est fiable, moins l'humain est préparé à gérer ses défaillances. Ce paradoxe a quarante ans. Il n'a pas vieilli d'un jour.

La confiance aveugle n'est pas un défaut corrigible par la formation. Dans un autre contexte métier où l'automatisation a pris le pas, Parasuraman et Manzey le démontrent[4] : c'est un mécanisme attentionnel. Quand un système automatisé fonctionne en arrière-plan, l'attention se déplace. La vigilance chute — chez les novices comme chez les experts. L'entraînement ne suffit pas. La structure doit compenser ce que l'attention ne fera pas.

Le deskilling est déjà mesurable. En médecine, le taux de détection d'adénomes par des endoscopistes chute de 28,4% à 22,4% après une exposition routinière à l'IA — quand ils travaillent ensuite sans assistance[14]. En pathologie, des praticiens abandonnent des diagnostics initialement corrects face à des suggestions erronées d'un système IA[15]. L'érosion est silencieuse, progressive, et elle touche les experts autant que les novices. Dreyfus[5] avait posé le cadre théorique — la progression vers la maîtrise passe par la pratique délibérée. Quand l'IA semble prendre en charge la difficulté, elle supprime l'occasion même de progresser. Les premières données empiriques vont dans ce sens.

## *La technologie a ses limites.*
## *La méthode aussi.*

Quand l'humain décroche, l'armature cède. Quatre modes de défaillance, tous documentés sur le terrain.

### CONTAMINATION FACTUELLE

**La contamination factuelle.**

Le web se contamine à grande échelle — les modèles s'entraînent sur leurs propres sorties, les erreurs se stabilisent, la correction devient impossible. Shumailov et al. appellent ça le model collapse[16] : chaque génération de modèle hérite des hallucinations de la précédente. Alemohammad et al. le formalisent.

À l'échelle d'un repo, le mécanisme est le même. Une donnée approximative entre une fois — parfois de l'humain, parfois hallucinée par l'IA — et se propage dans tous les documents générés ensuite. Sur Katen, ~30 documents contenaient "14 ans" au lieu de "18 ans" pour une durée de réflexion. L'erreur venait de l'humain lui-même, propagée et stabilisée par l'IA.

La différence : à l'échelle du web, c'est irréversible. Dans un repo SOFIA, c'est traçable et corrigeable. À condition que l'humain vérifie.

### VALIDATION SANS REGARD

**La validation sans regard.**

L'humain approuve sans lire, ou court-circuite la friction pour aller plus vite. Les sessions deviennent un rituel — ouvrir, valider, fermer. Un persona rédige et valide dans la même chaîne, sans challenger. C'est Bainbridge appliqué à la méthode : plus le système fonctionne bien, moins l'humain est vigilant.

Sur Katen, le produit (périmètre serré, production qualifiée) reste maîtrisé. Les explorations, elles, s'accumulent — documents non triés, productions non qualifiées, choses oubliées. La méthode qui fonctionne bien génère plus de matière que l'humain ne peut en absorber.

### DÉRIVE DE SCOPE

**La dérive de scope.**

Un persona mal recalibré absorbe le rôle des autres. Le calibrage initial ne suffit pas — les personas dérivent avec l'usage, et seul l'humain le voit.

Sur Katen, les périmètres de Sofia et Nora étaient définis par compétence métier — visual design, UX. Quand le livre bleu a dû exister en markdown, PDF, HTML et visuels réseaux, personne n'avait le contrat clair sur qui produit quelle transformation, pour quel canal. Les tâches tombaient entre les chaises — pas parce qu'un persona débordait, mais parce que la frontière était invisible.

### ANGLE MORT PARTAGÉ

**L'angle mort partagé.**

Les personas sont tous calibrés par le même humain. Ses biais implicites deviennent les biais de toute l'équipe. La friction est réelle — mais elle joue dans un espace de pensée borné par ce que l'orchestrateur sait qu'il ne sait pas. Ce qu'il ignore ignorer, aucun persona ne le lèvera.

Sur Katen, la méthode SOFIA documente bien la friction entre personas de réflexion — archi vs code vs stratégie. Elle ne documentait pas les chaînes de production multi-personas : qui publie quoi, sur quel canal, avec quel challenger. Aucun persona ne l'a levé spontanément. C'est la limite structurelle d'un système mono-orchestrateur — et la seule que la discipline ne résout pas. Elle demande un regard extérieur.

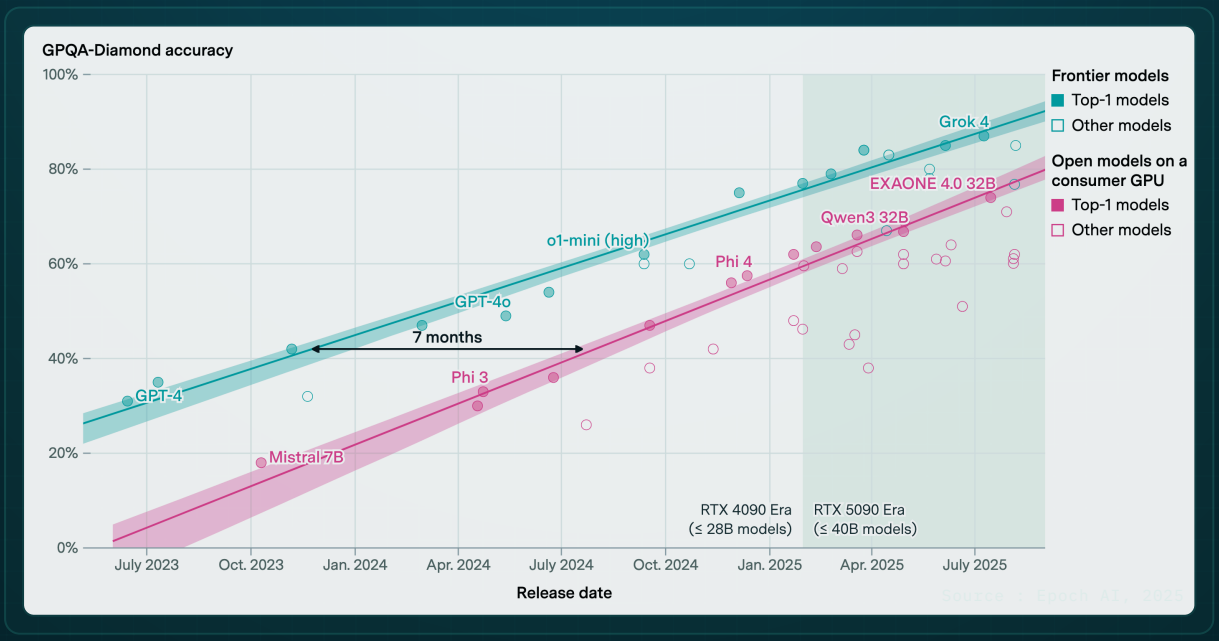
## Le transfert de charge

> **Le transfert de charge.** Ressources humaines constantes ne signifie pas ressources totales constantes. L'augmentation déplace une partie de l'effort vers l'infrastructure — tokens, compute, énergie. Il n'y a pas de magie : ce que l'humain ne porte plus, la machine le porte — et la machine consomme.

Sur Katen, une session de travail avec un persona consomme entre plusieurs milliers et plusieurs dizaines de milliers de tokens. Cinq personas, des sessions quotidiennes, des mois de projet — le volume cumulé est significatif. La méthode multiplie les interactions par design : friction, review croisée, itérations. Chaque passage de relai entre personas est un coût computationnel. L'approche full-auto consomme autrement — moins d'allers-retours, mais des chaînes plus longues et moins contrôlées. Le problème n'est pas propre à la méthode. Il est structurel à tout usage intensif de LLMs.

La méthode est soutenable pour l'humain. La question de sa soutenabilité énergétique à grande échelle reste ouverte — et l'honnêteté impose de ne pas l'esquiver.

Mais ce coût n'est pas statique. Epoch AI montre que les capacités des modèles frontier deviennent accessibles sur un GPU grand public en six à douze mois — et que l'écart se réduit[29]. Ce qui exige aujourd'hui une API facturée au token pourra demain tourner localement, sur du matériel à quelques milliers d'euros. Le transfert de charge reste réel, mais sa trajectoire est déflationniste.



> *Les outils full-auto ne documentent pas leurs modes de défaillance.*
> La méthode, si.

Les outils full-auto ne documentent pas leurs modes de défaillance. La méthode, si.

Ces limites ne disqualifient pas l'IA. Brynjolfsson et al. mesurent des gains de productivité à moyen terme sur des milliers d'agents de support[6]. Mais ces gains existent *sous conditions*. Elles cadrent ce qu'on peut en attendre — et ce qu'on ne doit pas lui déléguer sans filet.

Fragment VI

# Ce qui empêche l'armature de céder

*Les devoirs de l'orchestrateur*

La section précédente pose ce qui casse. Celle-ci pose ce qui tient.

Les risques de la méthode ne sont pas des fatalités. Ils se gouvernent — par des devoirs. Pas des recommandations. Des obligations que l'orchestrateur se donne et qu'il tient.

### 1 VÉRIFIER LES FAITS

**Vérifier les faits.** Le repo n'est pas une source de vérité pour les faits. La cohérence linguistique d'un LLM ne garantit pas sa véracité factuelle[21]. Une date approximative entrée une fois sera propagée dans tous les documents générés ensuite. Dates, durées, chiffres, noms propres, références : vérification humaine systématique, en continu — pas en fin de projet. Zhang et al. montrent que les LLMs peuvent identifier leurs propres hallucinations — une piste, pas une garantie[17].

### 2 ARBITRER

**Arbitrer.** Les personas exposent les tensions, ils ne les résolvent pas. Deux personas qui se contredisent indéfiniment ne produisent rien. L'humain écoute, questionne, puis tranche. La décision est tracée. Les personas appliquent, même s'ils avaient une position différente[22].

### 3 RELIRE CE QUI SORT

**Relire ce qui sort.** L'IA produit. L'humain publie. Entre les deux, il y a une relecture qui n'est pas de la correction — c'est de la validation. Aucun document ne sort du repo sans que l'humain l'ait lu en entier. Pas survolé, pas approuvé sur la base du résumé. Lu.

### 4 CALIBRER LES PERSONAS

**Calibrer les personas.** Un persona se définit par ce qu'il ne fait pas avant ce qu'il fait. Mais les bonnes contraintes émergent de l'usage. Un persona trop large dérive — il fait le travail des autres. Un persona trop étroit est inutile — personne ne lui parle. Le calibrage est continu. Il ne s'arrête pas après le bootstrapping.

### 5 SÉPARER RÉFLEXION ET PRODUCTION

**Séparer réflexion et production.** Un persona qui réfléchit et produit le livrable final est juge et partie. La friction disparaît parce qu'il n'y a personne pour challenger. Celui qui rédige n'est pas celui qui valide. La chaîne comporte au moins un regard extérieur avant la sortie.

### 6 MAINTENIR L'ATTENTION

**Maintenir l'attention.** Les personas ne détectent pas quand l'humain décroche — ils continuent à produire. Les signaux : tu approuves sans lire, les sessions deviennent mécaniques. L'automatisation ne supprime pas les difficultés, elle peut en créer de plus grandes[3], silencieusement. Quand tu reconnais ces signaux, c'est le moment de ralentir — pas d'accélérer. Et la friction qui protège le mieux est aussi celle qui résiste le plus — les utilisateurs la rejettent subjectivement[23].

# Le prix à payer

Six devoirs. Le prix à payer pour que l'armature tienne.

Ce document est lui-même un produit de la méthode qu'il décrit. Winston rédige. Mira dépose une review structurelle : *"Doublons structurels A, B, C — sévérité forte. Les sections débordent de leur rôle."* Léa dépose une review scientifique : *"[3] mélange Endsley (1995) et Huang (2025). Le lecteur académique pourrait croire que les chiffres datent de 1995. Séparer."* Marc dépose une review stratégique : *"Le chapitre II sonne comme 'il faut être moi pour que ça marche'. Ancrer sur l'expertise domaine comme mécanisme reproductible."* Trois reviews. Trois angles que les autres ne voyaient pas. Vingt points dont trois corrections factuelles, deux restructurations, un recadrage de posture. Le document est meilleur de trois façons indépendantes. Aucune ne serait venue d'un agent seul.

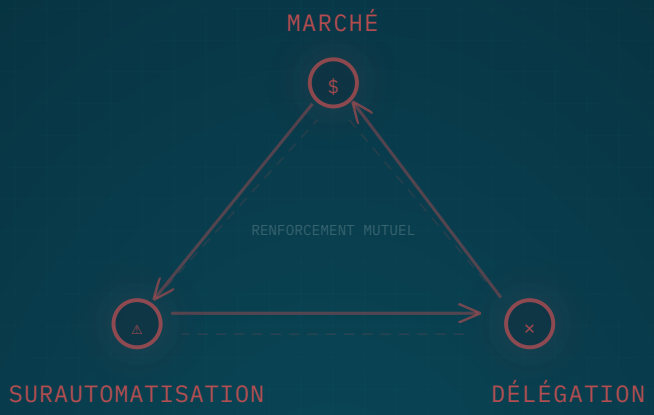
C'est une charge cognitive réelle. Six disciplines à tenir en parallèle, en continu, sans relâche. L'augmentation ne réduit pas l'effort — elle le déplace. Moins d'exécution, plus de vigilance. C'est le coût de la qualité. Et c'est exactement pour ça que l'humain n'est pas optionnel.

Fragment VII

# Le triangle de la dépendance

*Au-delà du projet*

Tout ce qui précède concerne un projet, une équipe, une méthode. Mais les enjeux ne s'arrêtent pas à l'échelle d'une organisation. Ils ont trois conséquences systémiques, et elles se renforcent mutuellement.




**La dépendance au bon marché.** L'IA est bon marché. Aujourd'hui. Le prix d'un token est une décision commerciale, pas une constante physique. Des organisations entières restructurent leur productivité autour de ce prix. C'est un pari sur la roadmap de quelqu'un d'autre. Le jour où le pricing change — et il changera — celles qui ont bâti leur fonctionnement dessus découvriront qu'elles ne savent plus fonctionner sans.

**L'éducation sans résilience.** Former avec l'IA sans apprendre à travailler sans, c'est produire une génération qui ne passera jamais le stade 3 de Dreyfus[5]. Le jugement se forme en affrontant la difficulté, pas en la contournant. Le vrai enjeu n'est pas "faut-il interdire l'IA en formation" — c'est à quel stade l'introduire, et avec quelle discipline.

**La perte de maîtrise.** La surautomatisation ne produit pas seulement de l'inefficacité. Elle produit de la dépendance stratégique. On a déjà vu ce film avec le cloud — la Cour des comptes a documenté comment dix ans de migration insuffisamment questionnée ont créé une dépendance structurelle de l'État français aux hyperscalers américains, transformant un choix technique en enjeu de souveraineté[24]. Le même schéma se reproduit avec l'IA, en plus rapide, en plus profond. Parce que la dépendance n'est pas seulement technique. Elle est cognitive.

## Le triangle n'est pas une fatalité.

*C'est un choix qu'on fait — ou qu'on laisse se faire.*

Les trois faces se renforcent. Le bon marché accélère la surautomatisation. La surautomatisation érode les compétences, ce qui rend l'éducation critique. Mais si l'éducation elle-même repose sur une IA bon marché sans discipline, elle produit des gens qui ne savent plus travailler sans. La boucle se ferme.

Dans les trois cas, le problème n'est pas l'IA. C'est l'absence de discipline dans son adoption.

Le triangle n'est pas une fatalité. C'est un choix qu'on fait — ou qu'on laisse se faire.

Fragment VIII

# De conviction, pas technique

*Le choix*

On mesure deux choses : la puissance brute des modèles et la durée d'autonomie des agents. On ne mesure pas la troisième — celle qui compte : où l'IA crée de la valeur, métier par métier, tâche par tâche.

Le benchmark unifié par verticale n'existe pas. Des benchmarks de niche existent — LegalBench, FinBen, MedQA, HumanEval[25] — mais personne ne les assemble en une carte cohérente. L'absence de mesure par verticale permet le récit "l'IA améliore tout" — qui justifie le déploiement massif — qui rend la mesure encore plus urgente.

L'angle mort n'est pas technique. Il est structurel. Les labos n'ont pas d'intérêt à prouver que l'IA ne marche pas dans un secteur. Les cabinets de conseil vendent du déploiement, pas du diagnostic. Les entreprises n'ont ni les outils ni la culture pour mesurer le gain réel.

Pendant ce temps, le déploiement n'attend pas la mesure. On automatise tous les métiers en parallèle, à l'aveugle.

Il y a deux écoles.

L'école full-auto dit : l'IA pense, l'humain valide. L'objectif — minimiser la friction, maximiser l'autonomie de l'agent. Le KPI — temps économisé, tâches complétées sans intervention. À l'échelle, la capacité de jugement se concentre chez ceux qui configurent les agents. Les autres deviennent opérateurs de systèmes qu'ils ne comprennent pas.

L'école humain augmenté dit : l'humain pense mieux grâce à l'IA. L'objectif — générer de la friction utile, amplifier le jugement. Le KPI — qualité des décisions, profondeur de la réflexion. L'humain reste orchestrateur, arbitre, penseur au centre.

Ce ne sont pas deux points sur un spectre. Ce sont deux philosophies incompatibles.

La première école est en train de gagner par défaut. Pas par supériorité — par visibilité. Les outils, le marketing, les benchmarks, les investissements vont tous dans cette direction. L'école full-auto scale. L'école augmentée demande une posture individuelle, une discipline, presque une éthique. Plus dur à industrialiser.

Mais quand tout le monde court dans la même direction, la question n'est pas "pourquoi tu ne suis pas ?" — c'est "où est-ce qu'ils vont tous, et qu'est-ce qu'il y a au bout ?"

Au bout de la surautomatisation : des systèmes que personne ne comprend, des compétences qui se sont évaporées, et un jour de panne où il n'y a plus personne pour réparer.

Au bout de l'augmentation : des gens qui pensent mieux, des systèmes qu'on maîtrise, et une technologie qui sert au lieu de remplacer.

---

*Le choix n'est pas technique. Il est de conviction.*

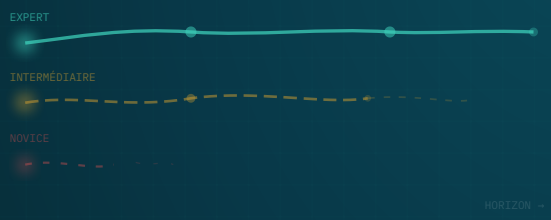
---

Fragment IX

# Ce que la méthode ne sait pas encore

*Travaux futurs*

*L'honnêteté d'une méthode se mesure à ce qu'elle admet ne pas savoir.*




## Trois profils, trois réalités

La méthode SOFIA suppose un orchestrateur avec une expertise domaine profonde. Cette condition n'est pas un détail — c'est le socle. Ce qui change selon le profil :

**Praticien expert** — plein régime. Les personas challengent, l'humain tranche avec discernement. La friction produit de la valeur parce que l'orchestrateur sait reconnaître quand l'IA se trompe.

**Praticien intermédiaire** — version contrainte. Moins de personas, friction réduite, validation externe obligatoire. La méthode peut servir de cadre d'apprentissage — mais elle ne remplace pas l'expérience qui manque. Le risque : croire que la friction suffit à compenser le manque de recul. Ce n'est pas le cas — la friction révèle les tensions, elle ne donne pas la compétence pour les résoudre.

**Novice** — la méthode ne s'applique pas. Plus grave : l'usage des LLMs sans expertise domaine est dangereux. Un novice ne détecte ni les erreurs ni les approximations. L'IA produit avec assurance, le novice valide sans recul. Ce n'est pas de l'assistance — c'est de la confusion stabilisée. L'illusion de compétence qu'installe un LLM bien formulé est plus nocive que l'absence d'outil — parce qu'elle supprime le signal d'alerte naturel qui pousse à chercher, vérifier, douter.

# Questions ouvertes

## MULTI-ORCHESTRATEUR

**Que devient la méthode quand plusieurs orchestrateurs partagent l'expertise ?**

SOFIA a été testée par un praticien seul qui porte la vision complète. À plusieurs, les dynamiques changent en profondeur. Qui arbitre quand les orchestrateurs ne sont pas d'accord ? Les personas répondent à une voix — si cette voix se divise, la cohérence des contraintes s'effrite. Chaque orchestrateur croit que l'autre couvre ce qu'il ne voit pas. Sans mécanisme explicite de synchronisation, les zones grises entre expertises deviennent les zones de moindre vigilance. Le risque est un consensus mou entre humains que les personas ne peuvent pas challenger.

## DÉGRADATION DE LA FRICTION

**Comment mesurer la dégradation de la friction dans le temps ?**

Les personas dérivent — leurs contraintes s'érodent à mesure que l'orchestrateur les recalibre vers le confort. L'humain décroche — la vigilance baisse quand le système fonctionne bien, exactement le paradoxe de Bainbridge décrit plus haut. La question n'est pas si ça arrive, mais quand. Et surtout : quel signal permet de le détecter avant que la friction ne devienne un rituel vide ?

## SEUIL DE PERSONAS

**À partir de combien de personas la charge cognitive devient-elle un goulot d'étranglement ?**

Cinq personas sur Katen — c'est gérable. Dix ? Quinze ? Il existe un seuil au-delà duquel l'orchestrateur ne peut plus tenir la carte complète des tensions. Il commence à déléguer mentalement, à faire confiance à certains personas sans les challenger. Et un persona non challengé, c'est un LLM en roue libre — exactement ce que la méthode cherche à empêcher.

## ANGLE MORT PARTAGÉ

**L'angle mort partagé.**

Quand l'orchestrateur et les personas partagent le même aveuglement — biais de domaine, culture technique, hypothèses implicites — aucune friction interne ne peut le révéler. La méthode n'a pas aujourd'hui de mécanisme de regard extérieur. C'est sa limite structurelle la plus sérieuse.

## TRANSFERT DE DOMAINE

**La méthode se transfère-t-elle à d'autres domaines ?**

Sur Katen, la friction a été testée en développement, en recherche et en design. Elle fonctionne — mais le terrain technique reste le plus naturel. Dans les domaines où les critères de qualité sont moins formalisables, la friction entre personas risque de tourner en boucle sans critère d'arbitrage clair. La question reste ouverte : qu'est-ce qui doit être adapté pour que la méthode tienne en dehors de son terrain d'origine ?

## MESURER LA VALEUR

**Comment mesurer la valeur produite par la friction ?**

La méthode repose sur un constat qualitatif : la friction améliore les artefacts. Mais on n'a pas de métrique. Comment comparer objectivement un livrable produit avec friction et le même sans ? Sans mesure, la démonstration repose sur la conviction du praticien et sur des exemples. C'est honnête — mais c'est insuffisant pour convaincre au-delà du cercle de ceux qui l'ont vécu.

SOFIA est ouverte. Le repo est là.
Le reste, c'est de l'expérimentation.

*La méthode est née d'un terrain — un projet, une personne. Chez toi, ce sera différent. Le format est libre — notes, reviews, ADR, conversations, ce qui colle à ta manière de travailler. La seule constante : pose des questions, challenge ce que les personas produisent, ne valide pas sans comprendre.*

La méthode ne marche que si tu la fais tienne.

`github.com/oxynoe-dev/sofia`

ANNEXE

# Lectures complémentaires

*Pour aller plus loin sur les fondations de ce livre bleu.*

### SUR L'AUTOMATISATION ET SES PARADOXES

Bainbridge, L. (1983). "Ironies of Automation." *Automatica* — Le texte fondateur. Quarante ans et pas une ride.

### SUR L'EXPERTISE ET L'APPRENTISSAGE

Dreyfus, H. L. & Dreyfus, S. E. (1986). *Mind Over Machine.* — Les cinq stades, de novice à expert. Pourquoi sauter des étapes ne fonctionne pas.

### SUR LA COLLABORATION HUMAIN-IA

Dell'Acqua, F. et al. (2023). "Navigating the Jagged Technological Frontier." *Organization Science* — 758 consultants BCG. Les gains existent, mais uniquement à l'intérieur de la frontière de compétence de l'IA.

### SUR LA CONTAMINATION DES DONNÉES

Shumailov, I. et al. (2024). "AI Models Collapse When Trained on Recursively Generated Data." *Nature* — Le model collapse, documenté.

Zhang, M. et al. (2024). "How Language Model Hallucinations Can Snowball." *ICML 2024* — Les LLMs peuvent identifier leurs propres hallucinations — une piste de mitigation.

### SUR LES SYSTÈMES MULTI-AGENTS ET LEURS DÉFAILLANCES

Cemri, M. et al. (2025). "Why Do Multi-Agent LLM Systems Fail?" *arXiv:2503.13657* — Taxonomie de 14 modes de défaillance sur 1642 traces d'exécution.

Huang, J.-T. et al. (2025). "On the Resilience of LLM-Based Multi-Agent Collaboration with Faulty Agents." *ICML 2025* — L'orchestrateur central est la topologie la plus robuste.

### SUR L'ARCHITECTURE DES AGENTS IA

Horthy, D. (2025). "12-Factor Agents — Principles for Building Reliable LLM Applications." *HumanLayer* — Douze principes d'ingénierie pour les agents LLM en production. Converge avec SOFIA sur les agents spécialisés (factor 10), la reprise du contrôle sur le flux (factor 8) et la transparence des prompts (factor 2). Diverge sur la friction : Dex optimise la fiabilité, SOFIA en fait un levier de qualité.[26]

Garcia, O. (2025). "Claude Buddy v5." *Plugin Claude Code, MIT* — 12 personas spécialisées activées à la demande par phase du cycle de dev (spec, plan, tasks, impl, docs), avec mémoire persistante entre sessions. Converge avec SOFIA sur les rôles spécialisés et l'activation humaine. Diverge : pas de friction entre personas, pas d'interdits, pas d'arbitrage — les personas assistent sans se challenger.[27]

### SUR LE DESKILLING

Budzyń, K. et al. (2025). *The Lancet Gastroenterology & Hepatology* — Deskilling mesuré chez les endoscopistes après exposition routinière à l'IA.

# Notes


[1] LeCun, Y. (2022). "A Path Towards Autonomous Machine Intelligence." Version 0.9.2, OpenReview. LeCun argumente que les LLMs autorégressifs sont une impasse fondamentale : pas de modèle du monde, pas de planification hiérarchique, pas de raisonnement — la prédiction de token ne mène pas à l'intelligence. openreview.net/pdf?id=BZ5a1r-kVsf

[2] Bainbridge, L. (1983). "Ironies of Automation." *Automatica*, 19(6), 775-779. ckrybus.com/static/papers/Bainbridge_1983_Automatica.pdf

[3] Endsley, M. R. & Kiris, E. O. (1995). "The Out-of-the-Loop Performance Problem and Level of Control in Automation." *Human Factors*, 37(2), 381-394. www.researchgate.net/publication/238726310_The_Out-of-the-Loop_Performance_Problem_and_Level_of_Control_in_Automation

[4] Parasuraman, R. & Manzey, D. (2010). "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors*, 52(3), 381-410. www.researchgate.net/publication/47792928_Complacency_and_Bias_in_Human_Use_of_Automation_An_Attentional_Integration

[5] Dreyfus, H. L. & Dreyfus, S. E. (1986). *Mind Over Machine.* Basil Blackwell. L'expertise se construit par la pratique délibérée — l'automatisation qui supprime la pratique empêche la progression vers la maîtrise. ISBN: 0-631-15126-5.

[6] Brynjolfsson, E., Li, D. & Raymond, L. (2025). "Generative AI at Work." *The Quarterly Journal of Economics*, 140(2), 889-942. 5172 agents de support, gains de productivité à moyen terme. L'étude suggère un effet de cliquet : les cibles de performance sont relevées quand trop d'agents les atteignent, ce qui peut absorber les gains. academic.oup.com/qje/article/140/2/889/7990658

[7] Forrester (2026). "The AI Layoff Trap: Why Half Will Be Quietly Rehired." *HR Executive*. Source commerciale — données non indépendantes, à corroborer. Cohérente avec les cas documentés [9][10][11] mais non vérifiable indépendamment. hrexecutive.com/the-ai-layoff-trap-why-half-will-be-quietly-rehired/

[8] Careerminds (2026). "The Cost of AI Layoffs." Enquête auprès de 600 HR leaders (fév. 2026) : 32,7% ont réembauché 25-50% des postes supprimés, 35,6% plus de la moitié. 30,9% ont dépensé plus en réembauche qu'ils n'avaient économisé. Source commerciale — données non indépendantes, à corroborer. careerminds.com/blog/cost-of-ai-layoffs

[9] Siemiatkowski, S. (2025). Interview Bloomberg, 8 mai 2025. *"It's so critical that you are clear to your customer that there will be always a human if you want."* Klarna relance l'embauche après l'effondrement qualité du chatbot IA déployé en 2024. www.bloomberg.com/news/articles/2025-05-08/klarna-turns-from-ai-to-real-person-customer-service — Voir aussi : www.usine-digitale.fr/article/klarna-retropedale-sur-l-ia-et-revient-aux-fondamentaux-du-service-client.N2231867

[10] IBM (2023-2025). Remplacement de ~8000 postes RH par le bot AskHR — 94% des requêtes automatisées, mais les 6% nécessitant empathie ou subjectivité ont conduit à une réembauche. Arvind Krishna, CEO (WSJ) : *"Our total employment has actually gone up, because what it does is it gives you more investment to put into other areas."* Signal convergent avec la thèse d'augmentation, mais déclaration d'un dirigeant intéressé. Sources secondaires multiples, pas de communiqué officiel IBM sur l'échec. www.hrkatha.com/news/ibm-rehires-after-ai-driven-layoffs-backfire-sparks-debate-on-automation-limits/

[11] McDonald's (2021-2024). Sources secondaires (CNBC, Fortune, CBS News), pas de communiqué officiel McDonald's sur les raisons de l'arrêt. IA vocale drive-through (partenariat IBM), déployée dans 100+ restaurants US. Précision visée : 95%+, obtenue : ~80% (BTIG, 2022). Commandes absurdes, accents incompris, corrections refusées. Programme arrêté juin 2024. www.cnbc.com/2024/06/17/mcdonalds-to-end-ibm-ai-drive-thru-test.html

[12] Krishnaprasad, M. (2025). Citation du CTO d'Agentforce dans *The Information* (déc. 2025) : "If you give an LLM more than, say, eight instructions, it kind of starts dropping instructions." Observation empirique de production, pas un seuil scientifique — la littérature montre une dégradation dépendante du modèle (loi de puissance). Contexte : Salesforce justifiait le passage d'Agentforce vers un raisonnement hybride LLM + logique déterministe. economictimes.indiatimes.com/news/new-updates/ai-bubble-bursting-salesforce-execs-admit-trust-issues-after-laying-off-4000-techies-now-scaling-back-use-of-ai-models/articleshow/126139465.cms

[13] Cemri, M., Pan, M. Z., Yang, S. et al. (2025). "Why Do Multi-Agent LLM Systems Fail?" *arXiv:2503.13657*. Preprint non peer-reviewed (auteurs issus de Berkeley). 1642 traces d'exécution, 7 frameworks (dont ChatDev, MetaGPT, AG2), 14 modes de défaillance classifiés en 3 catégories. Étude taxonomique — les auteurs classifient les causes d'échec, pas les taux de succès. arxiv.org/abs/2503.13657

[14] Budzyń, K. et al. (2025). "Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy." *The Lancet Gastroenterology & Hepatology*, 10(10), 896-903. Étude observationnelle rétrospective (4 centres polonais). Taux de détection d'adénomes : 28,4% → 22,4% après exposition routinière à l'IA. Les auteurs appellent à des essais confirmatoires. www.thelancet.com/journals/langas/article/PIIS2468-1253(25)00133-5/abstract — PubMed : pubmed.ncbi.nlm.nih.gov/40816301/

[15] Rosbach, E., Ganz, J., Ammeling, J., Riener, A. & Aubreville, M. (2024). "Automation Bias in AI-Assisted Medical Decision-Making under Time Pressure in Computational Pathology." *Springer* (2025). Preprint arXiv, petite étude (28 praticiens). 7% de diagnostics initialement corrects abandonnés face à des suggestions erronées d'un système IA. arxiv.org/abs/2411.00998

[16] Shumailov, I. et al. (2024). "AI Models Collapse When Trained on Recursively Generated Data." *Nature*, 631, 755-759. doi.org/10.1038/s41586-024-07566-y — Alemohammad, S. et al. (2024). "Self-Consuming Generative Models Go MAD." *ICLR 2024*. openreview.net/forum?id=ShjMHfmPs0

[17] Zhang, M. et al. (2024). "How Language Model Hallucinations Can Snowball." *ICML 2024*. Les LLMs peuvent identifier 67-94% de leurs propres claims incorrects quand évalués séparément — une piste de mitigation, pas une solution. proceedings.mlr.press/v235/zhang24ay.html

[18] Huang, J.-T. et al. (2025). "On the Resilience of LLM-Based Multi-Agent Collaboration with Faulty Agents." *ICML 2025*. La topologie hiérarchique avec orchestrateur central est la plus robuste face aux agents défaillants : chute de 5,5% vs 10-24% pour les topologies plates. Cette étude porte sur la résilience des topologies agent-agent. Le transfert à une architecture humain-arbitre repose sur l'hypothèse que l'humain est au moins aussi robuste qu'un orchestrateur IA central — hypothèse plausible mais non testée dans ce cadre. arxiv.org/abs/2408.00989

[19] Bruchon, J.-F. (2014). *Analyse par microtomographie aux rayons X de l'effondrement capillaire dans les matériaux granulaires.* Thèse, Université Paris-Est. Utilise Caméléon pour l'interaction live avec des algorithmes d'analyse d'image. pastel.hal.science/tel-01124287/file/2014PEST1007.pdf

[20] Cugnon de Sevricourt, O. & Tariel, V. (2011). "Cameleon language Part 1: Processor." *arXiv:1110.4802*. Modèle d'exécution basé sur une extension des réseaux de Petri pour un langage dataflow graphique d'analyse d'image interactive. arxiv.org/abs/1110.4802

[21] Ji, Z., Lee, N., Frieske, R. et al. (2023). "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys*, 55(12), art. 248. Taxonomie exhaustive : hallucinations intrinsèques (le texte se contredit lui-même) et extrinsèques (le texte contredit la réalité). La fluence et la cohérence masquent les erreurs factuelles — un texte peut sonner juste et être faux. doi.org/10.1145/3571730

[22] Shneiderman, B. (2020). "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy." *International Journal of Human-Computer Interaction*, 36(6), 495-504. Cadre HCAI : haute automatisation et haut contrôle humain coexistent — c'est une question de design, pas un compromis. L'humain conserve l'autorité de décision finale. arxiv.org/abs/2002.04087

[23] Buçinca, Z., Malaya, M. B., Gajos, K. Z. (2021). "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making." *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), art. 188. Sans friction délibérée, les humains acceptent les sorties IA sans les examiner. arxiv.org/abs/2102.09692

[24] Cour des comptes (2025). *Les enjeux de souveraineté des systèmes d'information civils de l'État.* Dix ans de migration cloud non questionnée ont créé une dépendance structurelle aux hyperscalers américains (~70% du marché cloud européen). www.ccomptes.fr/fr/publications/les-enjeux-de-souverainete-des-systemes-dinformation-civils-de-letat

[25] Benchmarks par verticale : LegalBench — Guha, N. et al. (2023). "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models." *NeurIPS 2023*. arxiv.org/abs/2308.11462 — FinBen — Xie, Q. et al. (2024). "FinBen: A Holistic Financial Benchmark for Large Language Models." *NeurIPS 2024*. arxiv.org/abs/2402.12659 — MedQA — Jin, D. et al. (2021). QCM basés sur USMLE. github.com/jind11/MedQA — HumanEval — Chen, M. et al. (2021). "Evaluating Large Language Models Trained on Code." OpenAI. github.com/openai/human-eval

[26] Horthy, D. (2025). "12-Factor Agents — Principles for Building Reliable LLM Applications." HumanLayer. Manifeste open-source (CC BY-SA 4.0) issu de l'accompagnement de dizaines de fondateurs YC. Les factors 2, 8 et 10 valident indépendamment des choix structurants de SOFIA — mais dans un cadre d'optimisation technique, sans la dimension friction/gouvernance. github.com/humanlayer/12-factor-agents

[27] Garcia, O. (2025). "Claude Buddy v5 — PAI-native development workflow platform." Plugin Claude Code open source (MIT). 7 skills, 12 personas, mémoire persistante. Convergence sur les rôles spécialisés et l'activation à la demande par l'humain. Absence de friction inter-personas et de gouvernance — les personas ne se challengent jamais, pas de mécanisme d'arbitrage. claude-buddy.dev/

[28] La Rosa, A. & Beretta, A. (2025). "Frictional AI in Joint Cognitive Systems: Towards a Human-Centered Approach at Higher Levels." *HHAI-WS 2025*, Pisa. CEUR Workshop Proceedings, Vol. 4074. Position paper (workshop short paper, pas article de recherche complet) — ouvre un agenda sur la friction comme élément de design dans les systèmes cognitifs conjoints multi-acteurs. Pas de validation empirique, mais première tentative de formalisation académique du passage de la dyade humain-IA au multi-acteurs. ceur-ws.org/Vol-4074/short3-1.pdf

[29] Somala, V. & Emberson, L. (2025). "Frontier AI capabilities can be run at home within a year or less." Epoch AI, août 2025. Les modèles open tournant sur un GPU grand public rattrapent les frontier en 6-12 mois, avec un écart qui se réduit (+125 Elo/an vs +80 Elo/an sur LM Arena). epoch.ai/data-insights/consumer-gpu-model-gap