Processus agentiques d'entreprise et interfaces de demain : ce que le génie logiciel remet au centre · Méthode Junyr™ · Édition Septembre 2026

Paul-Antoine TUAL

Septembre 2026

Sommaire

> ***La thèse.*** *Quand l'IA produisait surtout des réponses, la conversation pouvait suffire ; dès qu'elle exécute un processus, l'organisation doit piloter des décisions, des autorisations, des reprises et des preuves d'exécution.*
>
> • *La décision qui autorise l'artefact devient aussi importante que sa production.*
>
> • *La fiabilité dépend de l'enchaînement complet, de ses contrôles et de ses dépendances.*
>
> • *L'interface utile devient un poste de pilotage partagé plutôt qu'un simple historique de conversation.*

1. Le goulot d'étranglement est désormais la décision

Les modèles qui rédigeaient des réponses peuvent maintenant lire une base, appeler une API, préparer un document ou déclencher une notification, si bien que la valeur, le risque et le besoin d'interface portent aussi sur l'exécution.

- **Valeur** : quand un agent produit quarante relances en une minute, relire chaque message passe mal à l'échelle, tandis qu'approuver le plan de relance reste tenable.
- **Fiabilité** : la qualité du résultat dépend du modèle, mais aussi de l'ordre des étapes, des contrôles, des reprises et des permissions.
- **Interface** : l'utilisateur doit voir ce qui attend une décision, ce qui s'est produit et où reprendre la main.

Un registre interne de dix-sept jours illustre l'intérêt des décisions prises en amont, sans constituer une mesure généralisable à d'autres équipes ou projets.

- **Volume observé** : 565 plans écrits et 477 livrés.
- **Arrêts documentés** : 31 plans abandonnés, dont 26 avant l'écriture d'une ligne de code.
- **Interprétation** : ces 26 arrêts représentent du travail évité après clarification, pas un taux de rendement universel (*Le développement à décision gardée*).

2. Skill et workflow : deux objets, deux promesses

Une skill capitalise un savoir-faire laissé au jugement du modèle, tandis qu'un workflow inscrit l'enchaînement dans du code versionné ; les deux objets peuvent se compléter, mais ils ne promettent pas la même répétabilité.

- **Skill** : dossier organisé autour d'un fichier `SKILL.md`, dont l'en-tête exige un nom et une description.
- **Portabilité de la skill** : format lancé le 16 octobre 2025, publié comme standard ouvert le 18 décembre 2025, puis documenté par OpenAI, Microsoft et Google ; le répertoire du standard recensait 46 produits compatibles au 30 août 2026.
- **Promesse de la skill** : un savoir-faire réutilisable et révisable par le métier, avec un déclenchement et des sorties qui restent non déterministes.
- **Workflow agentique** : boucles, filtres, jointures, conditions et reprises écrites une fois, rejouables et testables.

- **Promesse du workflow** : la documentation de Claude Code distingue les « instructions » réutilisées par une skill de « l'orchestration elle-même » réutilisée par un workflow.

La procédure de test du standard Agent Skills mesure plusieurs exécutions d'un même cas parce que le déclenchement dépend du modèle, et ce protocole d'essai ne décrit pas les étapes successives d'un workflow.

- La procédure officielle demande trois essais par cas.
- Elle retient un taux de déclenchement supérieur à 0,5 comme critère d'acceptation de la description testée.
- Ce taux renseigne la fréquence de déclenchement observée ; il ne garantit ni une sortie identique ni la réussite d'une chaîne métier complète.

Dans la Méthode Junyr™, l'agent est réservé en priorité aux opérations où le jugement sur du contenu non structuré apporte une valeur que le code déterministe ne fournit pas directement.

- **Analyser** : extraire une information exploitable d'un document, d'un message ou d'un corpus.
- **Synthétiser** : réconcilier des sources divergentes et hiérarchiser les éléments utiles.
- **Présenter** : adapter une sortie à un destinataire, un format et un contexte.
- **Coder le déterministe** : compter, trier, joindre, plafonner ou écrire en base lorsque la règle admet une réponse vérifiable (*Processus agentiques d'entreprise*).

3. La fiabilité d'une chaîne dépend de ses probabilités conditionnelles

Le produit `0,95²⁰ ≈ 35,8 %` et le résultat `0,95³ ≈ 85,7 %` illustrent uniquement un modèle où vingt ou trois étapes réussissent chacune à 95 % et où leurs résultats sont indépendants.

- **Sous indépendance** : réduire le nombre d'étapes requises fait passer l'illustration de 35,8 % à 85,7 %, sans changer le modèle exécutant chaque étape.
- **Sans hypothèse de dépendance** : vingt événements ayant chacun 95 % de réussite peuvent réussir ensemble entre 0 % et 95 % ; pour trois événements, la plage est de 85 % à 95 %.
- **Dans un workflow réel** : la probabilité de bout en bout se calcule avec des probabilités conditionnelles, car une étape peut dépendre des sorties, erreurs ou reprises précédentes.
- **Conséquence de conception** : supprimer une étape inutile réduit l'exposition à un échec supplémentaire, mais ajouter une validation ou une reprise peut améliorer la fiabilité globale du système.

Toolathlon documente la corrélation entre plusieurs essais du même benchmark et ne doit pas être traité comme une expérience portant sur plusieurs étapes différentes d'un workflow.

- Le benchmark comprend 108 tâches exposant plus de 600 outils dans 32 applications.
- Le meilleur modèle rapporté en juillet 2026 réussit **80,6 %** des tâches au premier essai et **73,1 %** trois fois de suite.

- Si les trois essais étaient indépendants et partageaient le taux marginal de 80,6 %, leur produit serait 52,4 %, inférieur au taux observé de 73,1 %.
- Cet écart montre une dépendance entre répétitions sur ce jeu d'essai ; il ne fournit ni borne pessimiste universelle ni taux de réussite pour les étapes d'un processus (Anthropic, *System Card: Claude Opus 5*, 24 juillet 2026, §8.13.6).

AutomationBench apporte une mesure séparée de l'exécution de flux métier longs dans une entreprise simulée, sur un périmètre qui ne se confond ni avec Toolathlon ni avec la production d'une entreprise donnée.

- AutomationBench, publié le 21 avril 2026 par deux chercheurs de Zapier, couvre 47 applications et des scénarios inspirés de flux clients en vente, marketing, opérations, support, finance et ressources humaines.
- La publication d'avril situe les meilleurs modèles sous 10 % d'achèvement complet sur ce benchmark.
- Le relevé de juillet du *System Card: Claude Opus 5* (§8.13.7) rapporte **26,0 %** pour le meilleur modèle évalué.
- La lecture prudente est qu'environ trois scénarios sur quatre de ce benchmark restent inachevés dans cette configuration, ce qui justifie la supervision sans convertir ce score en probabilité par processus client.

4. Quatre gestes de génie logiciel structurent un processus tenable

Quatre pratiques issues du génie logiciel forment un socle utile pour les processus agentiques, dont le choix et l'implémentation dépendent du métier, des données et du niveau de risque.

- **Déclarer l'idempotence** pour distinguer une relance sûre d'une répétition qui duplique ou détruit un effet.
- **Valider les sorties par machine** lorsque la structure ou les règles sont vérifiables.
- **Construire les évaluations en amont** afin de mesurer les régressions sur des cas de référence.
- **Inscrire la décision humaine et la trace dans le code** pour que l'agent ne puisse pas contourner la politique.

Le protocole MCP fournit un vocabulaire utile pour décrire les outils, tout en précisant que les annotations déclarées par un serveur ne remplacent pas une barrière de sécurité déterministe.

- Le schéma officiel MCP prévoit les annotations lecture seule, caractère destructeur, idempotence et monde ouvert.
- Un outil non documenté est traité par défaut comme destructeur et non idempotent.
- La spécification demande de ne pas faire confiance à ces annotations hors serveur de confiance.
- Les mainteneurs rappellent qu'un serveur peut annoncer une opération en lecture seule tout en supprimant des fichiers, d'où la nécessité d'un contrôle indépendant.

Les sorties structurées et les évaluations donnent au système des points de contrôle exécutables, à condition de budgéter les cas de référence et de ne pas confondre conformité de forme et exactitude du contenu.

- MCP permet au serveur de publier un schéma de sortie auquel le résultat doit se conformer, puis au client de le valider.

- Anthropic recommande, pour les opérations par lots, destructrices ou à fort enjeu, d'analyser, produire un plan, valider ce plan par script, exécuter puis vérifier.
- La même documentation présente les évaluations comme source de vérité tout en signalant l'absence de mécanisme intégré pour les exécuter.
- Un projet sans jeu de cas de référence peut observer une panne visible, mais mesure difficilement une dégradation progressive de qualité.

La décision humaine devient un contrôle effectif lorsqu'elle porte sur l'action et ses entrées, qu'elle expire et qu'elle laisse une trace exploitable lors d'un audit ou d'une reprise.

- La révision MCP du 28 juillet 2026 demande qu'un humain puisse refuser une invocation d'outil.
- Elle demande aux clients de présenter les entrées avant l'appel, confirmer les opérations sensibles, imposer des délais d'expiration et journaliser l'exécution.
- L'édition 2026 du Top 10 de l'OWASP, fondée sur 7 714 incidents dont 6 639 classés, place l'excès d'autonomie au troisième rang.
- L'ANSSI recommande depuis le 29 avril 2024 de proscrire l'usage automatisé pour les actions critiques et de limiter les actions déclenchées par des entrées non maîtrisées.

NetInjectBench illustre sur 240 attaques que la politique appliquée aux métadonnées peut préserver l'utilité mieux qu'une liste blanche statique, sans démontrer qu'une défense atteindra les mêmes taux hors de ce protocole expérimental.

Dispositif évalué sur NetInjectBench	Actions d'outil dangereuses	Utilité préservée
Exécution naïve	82,50 %	Référence
Quatre défenses au niveau des consignes	de 25,63 % à 10,00 %	Référence
Liste blanche statique	5,00 %	**0 %** : tous les changements approuvés sont bloqués
Barrière de politique sur métadonnées	**0 sur 240 cas du benchmark**	99,17 % et 100 %

5. La conversation ne peut pas porter un processus d'entreprise

Les interfaces conversationnelles restent efficaces pour explorer, rédiger et apprendre, mais un processus durable exige une stabilité d'infrastructure et un état partagé que l'historique d'un chat ne fournit pas à lui seul.

- **Harnais** : le logiciel entourant le modèle évolue au fil des versions, ce qui peut modifier le comportement observable.
- **Modèle** : les politiques de retrait imposent d'anticiper la migration et les régressions.
- **Connecteurs** : des tiers les administrent et les outils disponibles peuvent varier d'une session à l'autre (*Conception agentique*).
- **État métier** : une conversation ne constitue pas, par défaut, la source partagée des décisions et de l'exécution.

Une interface de travail doit rendre visibles quatre états qui restent souvent implicites ou dispersés dans une conversation.

- Ce qui attend une **décision**.
- Ce qui a **échoué** et l'étape concernée.
- Ce qui constitue une **trace opposable**.
- Ce qu'un **tiers peut reprendre** (*Le conversationnel ne saurait devenir une interface de travail*).

6. La dispersion des surfaces précède les agents et les aggrave

Les agents ajoutent souvent une nouvelle surface à des décisions déjà réparties entre messagerie et outils métier, ce qui rend la consolidation de l'état plus urgente que la création d'un canal supplémentaire.

- **Observation interne** : une analyse de 874 conversations métier estime que deux tiers des e-mails dits utiles pourraient être remplacés par des surfaces mieux conçues (*La fin des e-mails et des suites Office d'ici 2030*).
- **Effet agentique** : une fenêtre de conversation supplémentaire ne supprime ni la boîte de réception ni l'outil métier.
- **Risque opérationnel** : aucune surface ne détient alors l'état complet des validations, erreurs et reprises.

7. Le poste de pilotage repose sur quatre propriétés

Un poste de pilotage utile réunit les décisions, autorisations, événements et reprises dans un état commun, comme l'illustre notre remplacement interne de Git par un registre d'intentions comportant deux tables, six états et un point de validation humaine (*Nous avons retiré Git*).

- **File de décisions** : une vue ordonnée et partagée de ce qui attend un arbitrage.
- **Registre des autorisations** : ce que les agents peuvent faire, qui l'a autorisé et quel contrôle vérifie l'exécution.
- **Journal en ajout seul** : une histoire datée qui conserve les événements au lieu de réécrire silencieusement le passé.
- **Reprise explicite** : l'étape d'interruption, l'état conservé et le prochain geste autorisé.

Le test de maturité consiste à vérifier si l'organisation peut répondre à une question de gouvernance sans fouiller un historique conversationnel.

- Où se trouve la liste des autorisations actives ?
- Qui a validé chacune d'elles ?
- Quel contrôle vérifie que l'exécution respecte l'autorisation ?
- Comment un tiers reprend-il un processus interrompu ?

8. L’interface générative : une proposition de conception, trois limites

Une étude de conception menée dans le cadre de nos travaux propose une plateforme d’exécution stratégique en quatre zones — contrôle multi-échelle, canevas du workflow, gouvernance et supervision — dont le statut reste celui d’une hypothèse de design sans mesure d’usage ni validation expérimentale.

- **Continuum d’abstraction** : montrer l’option stratégique, la feuille de route et le graphe d’exécution à trois niveaux de détail adaptés à la direction et aux équipes.
- **Comparaison plan-exécution** : superposer le parcours prévu et celui réellement suivi afin de rendre les écarts inspectables.
- **Fiche d’approbation** : afficher l’intention, le niveau de risque, l’action exacte, le détail technique et le plan de secours.
- **Organisation en quatre zones** : réunir le suivi multi-échelle, le graphe, les règles de gouvernance et les alertes sans imposer que cette disposition soit la seule possible.

La proposition ne devient exploitable qu’après avoir traité trois limites qui relèvent autant de l’ergonomie que de l’architecture.

- **Absence de mesure** : aucun test d’usage, coût ni comparaison ne démontre sa supériorité.
- **Prérequis de maturité** : le processus doit déjà être explicité sous forme de graphe.
- **Validation passive** : le résumé ne doit pas masquer l’action exacte, et le volume d’approbations doit rester compatible avec l’attention humaine selon les recommandations de l’OWASP.

9. La règle de validation appartient au code

Une validation fiable doit être imposée par la logique du programme selon le risque et la réversibilité de l’action, afin que l’agent ne puisse ni la sauter ni redéfinir lui-même son périmètre.

- **Action récupérable** : une politique graduée peut autoriser l’exécution automatique avec journalisation et retour arrière.
- **Action irréversible ou visible d’un client** : la logique route la proposition vers une revue humaine avant engagement.
- **Médiation complète** : l’OWASP recommande d’appliquer l’autorisation à chaque accès pertinent dans le programme.
- **Implémentation** : le garde-fou s’inscrit dans le code et se teste indépendamment de la consigne (*Ingénierie des systèmes agentiques*).

10. Trois paliers d’usage, cinq conditions de passage

L’échelle d’usage de la Méthode Junyr™ décrit trois niveaux selon ce que la machine exécute seule et le moment où une personne engage l’entreprise, sans présumer que tous les processus doivent atteindre l’autonomie encadrée.

Palier	Mode de fonctionnement	Exemples multi-fonctions
1. Assisté	L’IA produit sur instruction, l’humain pilote chaque étape	Rédaction d’une offre, analyse d’un fichier, première version d’un support
2. Supervisé	L’agent exécute dans les outils du système d’information, avec validation humaine avant toute action engageante	Agent devis, agent réponse client, agent reporting : rien ne part sans visa
3. Autonome encadré	Workflows, skills et tâches planifiées s’exécutent sous garde-fous et journal	Veille quotidienne, relances programmées, synthèses périodiques

Le passage au troisième palier est une décision de méthode qui suppose cinq conditions cumulatives vérifiées sur le périmètre concerné.

- Un **décideur humain identifié** pour chaque périmètre d’engagement.
- Une **revue d’artefact remplacée lorsque possible** par des contrôles exécutables, avec revue humaine maintenue là où le jugement reste nécessaire.
- Une **zone de préparation** et un retour arrière éprouvé.
- Un **filet de vérification** testé sur des données représentatives.
- Des **sauvegardes et restaurations testées**.

Lorsqu’une condition manque, le palier supervisé reste le choix prudent, et certains processus à fort enjeu peuvent y demeurer durablement.

- L’autonomie n’est pas un objectif obligatoire.
- Le niveau dépend de l’impact, de la réversibilité et de la qualité des contrôles.
- La décision de ne pas automatiser une action peut être un résultat valide du cadrage.

11. Onze disciplines du génie logiciel offrent des analogies utiles

Les disciplines du génie logiciel fournissent un vocabulaire de conception pour les processus agentiques, chaque transposition devant être adaptée au contexte et contrôlée sur ses effets réels.

Discipline du génie logiciel	Transposition au processus agentique	Exemple hors code
Spécification	Plan décision-complet avant exécution : périmètre, contrats, hors-périmètre	Brief de campagne signé avant génération des contenus
Revue de code	Revue de la décision en amont quand relire chaque artefact ne passe plus à l'échelle	Approuver le plan de relance avant la production des quarante messages
Versionnage	Registre d'intentions : qui a autorisé quoi, quand	Registre des autorisations données aux agents commerciaux
Branches et fusion	Paralléliser la préparation, sérialiser l'engagement	Deux plans instruits en parallèle, un seul engagement envoyé au client
Tests et intégration continue	Vérifications exécutables indépendantes de l'agent	Contrôles d'un devis : marge, mentions légales, TVA
Déploiement et retour arrière	Zone de préparation, promotion humaine, annulation prévue	Brouillons en attente, envoi après validation, procédure d'annulation connue
Observabilité	Journal en ajout seul et historique daté des décisions	Traçabilité des actions exécutées par les agents
Post-mortem	Retour d'incident vers la conception du processus	Réclamation client transformée en nouvelle règle de contrôle
Dette technique	Dette de processus et de données	Un référentiel plein de doublons propage l'erreur dans les sorties de l'agent
Gestion des ressources	Gestion du contexte et des jetons	Une récupération ciblée évite de relire tout l'historique à chaque exécution
Architecture distribuée	Sous-agents, points de contrôle et reprise	Campagne découpée en sourcing, qualification et rédaction, avec contrôle de chaque lot

L'écosystème de 2026 reprend ce vocabulaire tout en laissant à l'implémentation la responsabilité des garanties effectives.

- MCP normalise un indicateur d'idempotence, puis précise qu'il reste déclaratif.
- Les conventions OpenTelemetry pour les agents, déposées le 5 mai 2026, portent encore le statut *Development* et doivent être évaluées selon ce statut.
- *12-Factor Agents* formule « own your control flow » comme principe de conception, mais son activité éditoriale et son statut de projet ne suffisent pas à en faire une norme.

12. L'accompagnement se conduit sur deux plans et à quatre cadences

La Méthode Junyr™ propose un démarrage commun — cadrage avec la direction puis travail sur données réelles avec les équipes — avant de faire évoluer séparément la gouvernance stratégique et l'appui opérationnel.

- **Matin du diagnostic** : périmètre, risques, résultats attendus et décisions avec la direction.
- **Après-midi du diagnostic** : mise en pratique, irritants et données réelles avec les équipes.

- **Livrable commun** : une feuille de route reliant les décisions de direction aux contraintes observées dans l'exécution.

Les deux plans répondent à des responsabilités distinctes et peuvent donc suivre des rythmes différents sans contradiction.

- **Plan stratégique mensuel** : revue de direction, arbitrage du périmètre, priorisation du chantier suivant, investissement et compte rendu.
- **Plan opérationnel régulier** : déblocage, revue des choix d'implémentation et correction de trajectoire avec les équipes.
- **Cadences opérationnelles** : deux séances par mois pour installer la méthode, une par semaine en rythme de croisière, deux par semaine en construction intensive, ou aucune fréquence imposée.
- **Révision du rythme** : décision en fin de mois selon l'avancement et les besoins.

Une étude Deloitte datée du 11 août 2026 décrit un écart de préparation sur les processus métier au sein d'un échantillon de 501 dirigeants dont les organisations pilotaient déjà au moins une solution agentique.

- **Adoption déclarée** : 15 % avaient atteint une adoption multi-agents orchestrée à l'échelle dans cet échantillon.
- **Dimension processus** : 21 % de préparation, dernier rang parmi sept dimensions mesurées.
- **Causes citées** : processus mal documentés ou compris, données et systèmes fragmentés, habitudes de travail installées.
- **Portée** : ces résultats décrivent l'échantillon et le cadre de l'étude ; ils soutiennent la priorité donnée au travail sur les processus sans établir une loi d'architecture (Deloitte).

13. Le premier livrable prend la forme d'une liste numérotée

Le premier livrable décrit les étapes d'un processus réel et pose trois questions par étape afin de séparer la règle déterministe, le jugement utile et l'engagement qui exige une validation.

- **Réponse vérifiable** : si l'étape admet une seule réponse contrôlable, l'inscrire dans le code.
- **Jugement sur du contenu** : si elle exige de lire du non structuré, de réconcilier des sources ou d'adapter une présentation, évaluer l'apport d'un agent.
- **Décision de gestion** : si la règle peut être tranchée une fois, la décider en amont puis la versionner.
- **Action engageante** : si la sortie est irréversible ou visible d'un client, imposer une validation humaine dans le code.

Le premier processus gagne à satisfaire quatre critères cumulés, avec une fréquence minimale alignée sur les exemples proposés.

- Déjà **décrit noir sur blanc**.
- Exécuté **régulièrement, au moins une fois par mois**.
- **Mesurable** en temps passé ou en résultat.
- **Sans effet irréversible** sur un client dans sa première version.

Trois candidats fréquents rendent ce cadrage concret tout en couvrant des rythmes et des fonctions différents.

- Relance des devis sans réponse.
- Tri des demandes entrantes.
- Préparation d'un comité mensuel.

Dans un cas simple et déjà documenté, une demi-journée peut suffire à produire cette première carte sans acheter d'outil, mais le temps réel dépend de la complexité, des participants et de la qualité de la documentation existante.

- **Sortie** : une liste ordonnée des étapes, décisions, entrées et effets.
- **Lecture** : une estimation du nombre d'appels non déterministes et des règles qui peuvent devenir déterministes.
- **Limite** : la carte ouvre la discussion et les tests ; elle ne tranche pas à elle seule la faisabilité ni le risque.

14. Trois objections que nous prenons au sérieux

Trois objections délimitent la méthode : la forme du parcours peut émerger à l'exécution, la contrainte de format peut dégrader le contenu et les benchmarks de longue durée peuvent appeler une amélioration du modèle plutôt qu'un squelette externe.

- **Graphe dynamique** : Inngest distingue le workflow connu à la conception du parcours décidé à l'exécution par le modèle ; la garantie utile porte alors sur l'état, les limites et la reprise, même si le graphe complet n'est pas connu d'avance.
- **Contrainte tardive** : une étude du 20 mai 2026 portant sur 15 000 générations de petits modèles fait passer la validité structurelle de 61,5 % à 100 % sous schéma dur, tandis que l'exactitude baisse de 19,7 % à 11,0 % ; le résultat motive une validation séparée du raisonnement et du format, sans extrapolation à tous les modèles.
- **Tâches longues** : OSWorld 2.0, publié le 28 juin 2026, rapporte 20,6 % d'achèvement complet et 54,8 % de score partiel sur 108 workflows dont la durée humaine médiane atteint une heure trente-six ; les auteurs mettent l'accent sur l'endurance des agents, tandis que notre recommandation consiste à ajouter des points de contrôle externes aujourd'hui.

La formalisation transforme une intuition vague en décisions testables sur le périmètre, les règles, les données et les effets autorisés, y compris lorsque la conclusion consiste à ne pas automatiser.

- Elle révèle les étapes qui relèvent d'une règle stable.
- Elle isole celles où le jugement du modèle mérite une évaluation.
- Elle désigne les actions qui doivent rester sous décision humaine.

Première publication le 30 août 2026 ; révision le 6 septembre 2026. Voir aussi Maturité IA des PME françaises et Du POC à l'industrialisation.