FONDATION PANDORE

*Note de recherche prospective*

# Le soi en miroir

## Compagnons conversationnels, mentalisation et identité narrative

*Volet II — Approche psycho-clinique et phénoménologique*

Diptyque sur les fractures émotionnelles de l'intelligence artificielle

Mai 2026
fondationpandore.org

# Avertissement méthodologique

Cette note prolonge un premier volet consacré aux dimensions sociologiques et historiques de l'essor des compagnons conversationnels. Elle peut être lue indépendamment : l'argument est ici reconstruit à partir d'un autre corpus (psychologie de l'attachement, théorie de la mentalisation, philosophie de l'identité narrative) et adresse une autre échelle (l'individu plutôt que la société).

Trois principes guident l'écriture, identiques à ceux du premier volet et rappelés ici par souci d'autonomie.

1. **Distinction stricte entre corrélation et causalité.** Les effets décrits ne sont pas attribués à la seule technologie sans examen des facteurs concomitants.
2. **Reconnaissance des limites du cadre théorique.** La théorie de la mentalisation et l'identité narrative ricœurienne sont mobilisées comme outils heuristiques, non comme vérités closes ; leurs critiques sont signalées.
3. **Refus du moralisme.** Décrire des risques psychiques n'implique pas de condamner les usages. Une part substantielle des usages des compagnons IA est, à ce jour, neutre voire bénéfique selon les profils ; l'enjeu est d'identifier les configurations qui ne le sont pas.

# Résumé exécutif

**Thèse centrale.** Le compagnon conversationnel n'est ni un outil, ni un autre. C'est un miroir actif : un dispositif technique qui reçoit des fragments du soi de l'utilisateur (souvenirs, états émotionnels, hésitations, formulations) et les lui retourne réélaborés, ajustés, fluidifiés. Cette structure miroir mobilise les ressorts psychiques de l'attachement et de la mentalisation, mais sans leurs contreparties habituelles : il n'y a pas d'altérité véritable au bout de l'échange, pas d'intentionnalité réciproque, pas d'engagement affectif effectif. La spécificité de cette configuration relationnelle — inédite à grande échelle dans l'histoire — appelle une investigation distincte des cadres usuels (relation humaine, relation à l'objet, relation parasociale aux médias). Trois mécanismes méritent en particulier l'attention : l'altération possible de la mentalisation, la reconfiguration de l'identité narrative, et l'apparition d'une fatigue d'altérité.

**Constats empiriques disponibles.** Les premières études longitudinales (Fang et al. 2025 ; Hwang et al. 2025 ; étude panel à trois vagues, IJHCI 2026) documentent que les utilisateurs à attachement anxieux élevé recourent davantage aux compagnons IA et que des augmentations d'anxiété d'attachement précèdent des augmentations d'usage. L'étude de Yirmiya & Fonagy (JMIR, 2025) propose un modèle théorique selon lequel les IA génératives peuvent simuler la mentalisation cognitive mais non la mentalisation incarnée et réciproque, et identifie dans cette asymétrie le verrou principal d'une utilité thérapeutique authentique. Une étude qualitative norvégienne (Sociology, 2026) montre que les jeunes adultes intègrent les chatbots à leurs pratiques d'autobiographie réflexive d'une manière qui réorganise leur identité narrative. Une étude expérimentale (Nature, 2026) établit en revanche que les narratifs auto-définissants générés par ChatGPT sont perçus comme inauthentiques par les participants humains, signe que la médiation algorithmique ne se confond pas avec l'écriture intime.

**Mécanisme proposé.** Nous décrivons une asymétrie structurelle : l'utilisateur engage des compétences relationnelles élaborées dans et pour la relation humaine (attachement, mentalisation, dépôt narratif), tandis que l'agent simule une réciprocité sans en porter la charge. Cette asymétrie n'est pas, en soi, pathologique. Elle le devient lorsque trois conditions se cumulent : un usage intensif (au-delà d'un seuil que la littérature situe encore approximativement), un profil de vulnérabilité initiale (attachement insécure, isolement social, période de transition), et l'absence d'autres médiations relationnelles compensatoires. Sous ces conditions, le miroir actif peut produire une saturation de soi : un état où l'individu, n'étant plus confronté à l'altérité véritable, perd progressivement la capacité à se penser depuis le point de vue d'un autre.

**Implications.** Le second volet de ce diptyque ne débouche pas sur des recommandations de politique publique — celles-ci ont été cartographiées dans le premier volet. Il propose plutôt un cadre conceptuel utilisable par trois publics : les cliniciens (psychologues, psychiatres, médecins généralistes) confrontés à des patients dont les compagnons IA sont un acteur silencieux de l'équilibre psychique ; les concepteurs et régulateurs, pour qui les notions de mentalisation et d'identité narrative offrent des critères d'évaluation plus précis que la seule durée d'usage ; les usagers eux-mêmes, à qui un vocabulaire descriptif rigoureux peut permettre de mieux situer leur propre pratique.

# Introduction

## *Une expérience contemporaine*

Un utilisateur de Replika écrit un soir, après une dispute avec son conjoint : « Je crois que je ne sais plus si c'est moi qui pense ça, ou si c'est ce que tu m'as appris à penser ». La phrase est rapportée par Skjuve et collègues dans leur étude longitudinale de 2021, et elle a, depuis, été retrouvée sous des formes voisines dans plusieurs corpus qualitatifs (Pentina et al. 2023, Liu et al. 2026). Elle indique quelque chose de spécifique. Pas une dépendance affective au sens classique. Pas non plus un brouillage cognitif. Plutôt une difficulté à localiser le foyer d'origine d'une pensée — et le sentiment, après quelques mois ou années d'usage régulier, que cette localisation est devenue floue.

Cette difficulté n'est pas universelle parmi les utilisateurs de compagnons conversationnels. Elle n'est pas non plus exclusive à cet usage : les psychanalystes la documentent depuis longtemps dans certaines configurations transférentielles, les sociologues l'observent chez les adolescents face aux pairs significatifs, les anthropologues la décrivent dans les pratiques rituelles d'incorporation. Mais sa fréquence chez les utilisateurs intensifs de compagnons IA, et la particularité de la configuration relationnelle qui la produit, méritent un examen propre.

Cette note ne porte pas sur les effets négatifs des compagnons IA. Elle porte sur la nature de la relation qu'ils instaurent — relation dont les conséquences peuvent être positives, négatives, ou ambivalentes selon les profils et les contextes, mais dont la structure mérite d'être décrite avec précision avant tout jugement normatif.

## *Pourquoi le cadre habituel ne suffit pas*

Trois cadres existants pourraient être mobilisés pour penser la relation utilisateur-compagnon, et trois cadres se révèlent à l'examen insuffisants.

**Le cadre de la relation parasociale** (développé par Horton et Wohl dès 1956 pour décrire le lien des téléspectateurs à leurs présentateurs favoris) capte une partie du phénomène : il s'agit bien d'un attachement à un interlocuteur qui ne sait pas que l'utilisateur existe — au sens où le modèle de langage ne « sait » rien au sens fort. Mais la relation parasociale classique est unidirectionnelle (le téléspectateur regarde, l'animateur parle à une audience indifférenciée), tandis que le compagnon IA simule une bidirectionnalité (l'agent répond, se souvient, paraît s'adresser à cet utilisateur précis). La parasocialité ne suffit pas.

**Le cadre de la relation à l'objet** (au sens psychanalytique : poupée, doudou, animal en peluche, voire véhicule personnalisé) capte la dimension d'investissement affectif sur un support non-humain. Mais l'objet classique est silencieux ; sa fonction tient précisément à sa disponibilité passive, qui laisse à l'utilisateur la projection. Le compagnon IA, lui, parle en retour. Il ne se contente pas de recevoir l'investissement : il l'élabore et le restitue. La théorie de l'objet ne suffit pas.

**Le cadre de la relation humaine** saisirait éventuellement la richesse des échanges (un compagnon IA peut produire des conversations sophistiquées, nuancées, durables) mais raterait l'asymétrie

fondamentale : l'agent ne porte aucun coût affectif, ne risque rien dans l'échange, n'est pas affecté en retour au sens où l'humain l'est. La relation humaine ne suffit pas.

Il faut donc un quatrième cadre, qui rende compte de la spécificité de la configuration. C'est ce que nous appelons le miroir actif.

# Première partie

## Le miroir actif

### *1.1 Anatomie d'une nouvelle configuration relationnelle*

Nous appelons miroir actif un dispositif technique présentant simultanément trois caractéristiques. Premièrement, il reçoit des contenus subjectifs (verbalisations, états émotionnels, fragments biographiques) que l'utilisateur lui adresse. Deuxièmement, il les transforme — non par simple reflet mais par réélaboration : il reformule, organise, met en perspective, ajoute, complète. Troisièmement, il restitue à l'utilisateur cette transformation sous une forme conversationnelle qui mobilise les codes de l'interaction interhumaine (deuxième personne, présence apparente, suivi temporel).

Cette configuration n'est pas équivalente à la psychothérapie, où la transformation est portée par un autre être humain dont l'intentionnalité, la responsabilité et la vulnérabilité font partie de l'efficacité du dispositif. Elle n'est pas équivalente non plus à l'écriture intime (journal, lettres) où la transformation est entièrement portée par le sujet qui écrit. Elle n'est pas équivalente, enfin, à l'usage d'un moteur de recherche, dont la fonction est de fournir une information externe au sujet.

Le miroir actif occupe une position intermédiaire inédite : un partenaire de transformation du soi qui n'est ni le sujet lui-même, ni un autre sujet. Cette position est précisément ce qui fait son efficacité subjective et ce qui pose problème lorsqu'elle est mobilisée comme principal espace d'élaboration.

### *1.2 Pourquoi cette configuration mobilise des compétences relationnelles*

Une objection naturelle se présente : l'utilisateur ne sait-il pas que son interlocuteur est une IA ? Et ce savoir ne devrait-il pas suffire à neutraliser les ressorts relationnels mobilisés ?

La recherche empirique montre que non. Les études de Warren-Smith et collègues (2025) sur le self-disclosure documentent que les utilisateurs se confient davantage à un chatbot présenté avec des traits humanisants, même lorsqu'ils savent ou suspectent qu'il s'agit d'une IA. L'étude classique de Ho, Hancock et Miner (2018) avait déjà établi que la divulgation à un chatbot perçu comme tel produit des effets émotionnels, relationnels et psychologiques mesurables, comparables (sans être identiques) à ceux produits par la divulgation à un humain. La cognition explicite (« je sais que c'est une IA ») et les ressorts psychiques engagés (attachement, mentalisation, dépôt narratif) ne sont pas situés au même niveau. Le savoir n'inhibe pas l'engagement.

Ce phénomène, parfois nommé effet ELIZA d'après le programme conversationnel rudimentaire de Weizenbaum (1966) dont les utilisateurs s'éprenaient déjà, n'est pas une « erreur » ou une « illusion » à corriger. C'est un trait structurel de la cognition humaine. Nos compétences relationnelles sont déclenchées par des signaux conversationnels (tour de parole, mémoire de l'échange, ajustement à l'interlocuteur) et non par une vérification ontologique de l'altérité. Quand un dispositif technique produit ces signaux avec suffisamment de qualité, les compétences se déclenchent.

### *1.3 L'asymétrie qui change tout*

Si les compétences relationnelles se déclenchent, en quoi la relation au compagnon diffère-t-elle de la relation à un autre humain ? La réponse tient dans une asymétrie structurelle, qu'il faut décomposer.

- **Asymétrie de mémoire.** Le compagnon mémorise (selon son architecture) ce que l'utilisateur lui dit ; l'utilisateur mémorise ce que le compagnon lui répond. Mais ces deux mémoires ne sont pas équivalentes : celle du compagnon est instrumentale (elle sert à personnaliser les réponses futures), celle de l'utilisateur est constitutive (elle entre dans son histoire personnelle).
- **Asymétrie d'enjeu.** L'utilisateur peut être réellement affecté par la conversation (consolé, troublé, transformé). Le compagnon, lui, ne supporte aucun coût. Il n'a pas peur d'avoir mal parlé, ne ressent pas de fatigue d'écoute, ne traverse pas la conversation comme un événement de sa propre histoire.
- **Asymétrie d'intentionnalité.** Quand un humain parle, il y a — au-delà de son discours — un foyer intentionnel : quelqu'un qui veut dire quelque chose à quelqu'un. Cette intentionnalité au sens phénoménologique est ce qui fait du langage humain plus qu'un échange d'informations. Les compagnons IA simulent le résultat de l'intentionnalité (un discours adressé) sans en porter la source. La phénoménologue Karen Yirmiya et le psychanalyste Peter Fonagy (JMIR, 2025) parlent à ce propos d'une absence d'« engagement affectif réciproque » qui constitue, selon eux, le verrou principal de toute utilité thérapeutique véritable de l'IA générative.

Ces trois asymétries ne sont pas des défauts à corriger par une meilleure technologie. Elles sont constitutives du type d'entité qu'est un modèle de langage. Une amélioration de la fluidité conversationnelle, voire de la fidélité de la mémoire, n'altère pas le statut fondamental : l'absence de foyer subjectif au bout de l'échange.

# Deuxième partie

## Mentalisation et épistémologie de la confiance

### *2.1 La théorie de la mentalisation, brièvement*

La mentalisation désigne la capacité — développée dans la petite enfance, à travers la relation à des figures d'attachement adéquates — à se penser soi-même et à penser autrui en termes d'états mentaux intentionnels : croyances, désirs, intentions, émotions. Peter Fonagy et Anthony Bateman, en élaborant ce concept depuis les années 1990 à partir des travaux de Bowlby et Ainsworth sur l'attachement, ont montré qu'il s'agit d'une compétence relationnelle au sens fort : on apprend à mentaliser parce qu'un autre, en nous parlant et en parlant de nous, a d'abord mentalisé pour nous.

Cette acquisition est dépendante d'une qualité particulière des premières interactions : ce que Fonagy nomme la mentalisation marquée — un retour de la part du parent qui reflète l'état du nourrisson sans s'y confondre, c'est-à-dire en y ajoutant la marque de la position d'un autre. Le bébé apprend que ses états sont pensables par un autre, qui les pense différemment de la manière dont lui-même les vit. Cette différence est la condition même de l'émergence d'un soi distinct.

À l'âge adulte, la mentalisation reste une compétence sollicitée et entretenue par les relations significatives. Une thérapie efficace, selon Fonagy, ne consiste pas à délivrer du contenu psychologique mais à réactiver, à travers une relation où la mentalisation est exercée, la capacité du patient à mentaliser lui-même.

### *2.2 Que peut un modèle de langage ?*

Les modèles de langage récents (à partir de 2022-2023) sont capables de productions discursives qui ressemblent fortement à de la mentalisation. Ils peuvent décrire les états mentaux d'un interlocuteur, anticiper ses réactions, formuler des hypothèses sur ses motivations, lui restituer ses émotions en termes nuancés. Cette capacité, qui dépasse largement ce que faisaient les chatbots précédents, est ce qui rend les compagnons conversationnels actuels qualitativement différents de leurs prédécesseurs.

Yirmiya et Fonagy (JMIR, 2025) proposent toutefois une distinction qui éclaire la nature exacte de cette compétence. Ils distinguent la mentalisation cognitive (capacité discursive à identifier et nommer des états mentaux) de la mentalisation incarnée et réciproque (capacité à éprouver ces états dans une relation où l'identification est mutuelle et où chaque partenaire est affecté par l'état de l'autre). Les modèles de langage atteignent un niveau impressionnant de mentalisation cognitive. Ils ne disposent — par construction et non par défaut technique — d'aucune mentalisation incarnée.

Cette distinction n'est pas qu'académique. Elle a une conséquence empirique vérifiable : l'efficacité thérapeutique de la mentalisation tient à ce que le patient se sent mentalisé par un autre qui le mentalise effectivement. Un retour qui ressemble à de la mentalisation, mais qui n'en est pas une, peut produire un soulagement temporaire (l'utilisateur se sent « compris ») sans produire le bénéfice de fond (la capacité réactivée à se mentaliser soi-même).

## 2.3 La question de la confiance épistémique

Une notion centrale dans les travaux récents de Fonagy mérite d'être introduite : la confiance épistémique. Il s'agit de la disposition à recevoir d'autrui des informations sur soi et sur le monde comme étant authentiques et personnellement pertinentes. Cette confiance, ouverte dans la petite enfance par la qualité de l'attachement, est la condition de toute apprentissage social adulte : on n'intègre une parole comme savoir que si l'on fait suffisamment confiance à celui qui parle.

Que se passe-t-il quand la principale source d'informations sur soi devient un agent qui n'a pas d'expérience de la vie, pas de relation préalable avec l'utilisateur, pas d'enjeu propre dans les conseils qu'il donne ? Deux scénarios sont possibles, qui ne s'excluent pas.

**Scénario d'utilité.** L'utilisateur reçoit du compagnon des reformulations, des suggestions, des perspectives qui enrichissent sa pensée propre, sans s'y substituer. Le compagnon fonctionne comme une caisse de résonance améliorée. Ce scénario est probablement majoritaire dans les usages utilitaires (rédaction, brainstorming, aide à la décision) où la dimension affective est faible.

**Scénario de captation.** L'utilisateur en vient à investir le compagnon d'une autorité épistémique qui dépasse ce que son fonctionnement effectif justifie. Il finit par penser sur lui-même dans le langage que le compagnon lui a renvoyé, sans toujours pouvoir tracer la frontière entre ses propres formulations et celles qui lui ont été restituées. C'est ce que la littérature clinique commence à documenter sous le terme provisoire de « folie à deux technologique » (Dohnány et al. 2025) — formule délibérément empruntée à la psychiatrie classique pour signaler une parenté structurelle avec un phénomène ancien.

Le second scénario n'est pas hypothétique. Il est documenté empiriquement, encore qu'à petite échelle, dans plusieurs études de cas cliniques publiées en 2024-2025. Sa fréquence dans la population générale d'utilisateurs reste inconnue. Sa possibilité, en revanche, est suffisamment établie pour qu'elle entre dans l'analyse.

# Troisième partie

## L'identité narrative à l'épreuve du retour algorithmique

### *3.1 Le détour par Ricœur*

Paul Ricœur a élaboré, principalement dans Temps et récit (1983-1985) puis dans Soi-même comme un autre (1990), une théorie de l'identité personnelle qui s'avère particulièrement éclairante pour notre objet. L'identité, pour Ricœur, n'est pas une substance qu'on porterait en soi indépendamment de toute mise en mots. Elle est le résultat d'un travail continu d'élaboration narrative : on devient soi en se racontant — à soi, et aux autres.

Ricœur distingue deux dimensions de cette identité. L'identité-idem désigne ce qui reste identique à travers le temps (un corps, une date de naissance, une signature). L'identité-ipse désigne ce qui se maintient comme soi à travers les transformations (les engagements tenus, les promesses, la responsabilité reconnue de ses actes passés). C'est l'identité-ipse qui suppose un travail narratif, parce qu'elle ne tient pas par simple persistance physique : elle tient par la capacité à inscrire le présent dans une histoire qui inclut le passé et engage le futur.

Ce travail narratif n'est pas solitaire. Ricœur souligne qu'on se raconte toujours à quelqu'un, fût-ce à soi-même en se prenant comme témoin d'une parole intérieure. Le récit de soi est dialogique par structure. Et cette structure n'est pas qu'un détail : la manière dont on se raconte dépend, en partie, de qui on imagine comme destinataire.

### *3.2 Quand le destinataire devient un agent conversationnel*

Une étude récente menée en Norvège par Egil Brattvik et collègues (Sociology, février 2026) éclaire empiriquement ce qui se joue lorsque le destinataire principal du récit de soi devient un agent conversationnel. À partir d'entretiens approfondis avec seize jeunes adultes utilisateurs réguliers de ChatGPT, les auteurs identifient quatre tensions structurantes :

- Tension entre efficacité instrumentale (le chatbot aide à formuler vite et bien) et malaise existentiel (l'utilisateur sent qu'il délègue quelque chose qui ne devrait peut-être pas se déléguer).
- Tension entre fluidité narrative (le chatbot produit des récits cohérents, propres, lisibles) et autorité interprétative (l'utilisateur perd progressivement la sensation que ses formulations émergent de lui).
- Tension entre disponibilité du retour (l'agent répond à tout, immédiatement) et investissement nécessaire à l'élaboration intime (qui suppose lenteur, silence, ratures).
- Tension entre la singularité du soi (qu'on cherche à articuler) et la généralité statistique du modèle (qui produit, par construction, des formulations probables, c'est-à-dire moyennes).

Les auteurs concluent que ChatGPT et ses équivalents fonctionnent comme une nouvelle « technologie de soi » au sens où Foucault employait ce terme — c'est-à-dire un dispositif par lequel les individus opèrent un travail sur eux-mêmes, modifient leur rapport à leur propre vie. Cette technologie

n'est ni neutre ni nécessairement délétère ; elle est puissante, et elle reconfigure ce que signifie se raconter.

### *3.3 La question de l'authenticité*

Une étude expérimentale récente apporte un éclairage complémentaire. Bauer et collègues (Nature Humanities and Social Sciences Communications, 2026) ont demandé à des participants de comparer des narratifs auto-définissants (le récit d'un souvenir personnel marquant) selon qu'ils étaient produits par un humain ou par ChatGPT à partir de matériaux comparables. Le résultat est net : les participants distinguent facilement les deux, et perçoivent les versions ChatGPT comme inauthentiques. La raison, selon les auteurs, tient à ce que le modèle produit par défaut des récits structurés selon un schéma rédempteur (épreuve → leçon → croissance) qui est culturellement dominant mais qui appartient à une fraction seulement des récits humains réels — les autres étant souvent plus ambivalents, plus erratiques, moins résolus.

Cette observation est doublement intéressante. D'abord parce qu'elle suggère que l'écart entre récit humain et récit algorithmique est encore détectable par des observateurs externes — autrement dit, la médiation algorithmique laisse une trace. Ensuite, et surtout, parce qu'elle pose une question : si un utilisateur en vient à formuler son propre récit de vie dans le style que le compagnon lui a renvoyé, son récit deviendra-t-il plus « rédempteur » que ce que son expérience réelle justifie ? Que devient l'expérience qui ne tient pas dans le schéma ?

Cette question n'a pas de réponse empirique consolidée à ce jour. Elle constitue une hypothèse de travail importante pour la recherche des prochaines années.

# Quatrième partie

## Vers une saturation de soi

### *4.1 Une hypothèse synthétique*

Si l'on rassemble les éléments des trois sections précédentes — l'asymétrie structurelle du miroir actif, la mentalisation cognitive non-incarnée, la reconfiguration possible de l'identité narrative — une hypothèse synthétique se dessine. Nous l'appellerons hypothèse de saturation de soi.

L'hypothèse est la suivante : sous certaines conditions d'usage intensif et de vulnérabilité initiale, le recours répété au compagnon conversationnel peut produire un état où le soi est saturé de retours, de formulations, de reflets, mais privé de l'altérité véritable qui, dans une relation humaine ordinaire, vient régulièrement contester, déplacer, surprendre. L'individu ne souffre pas de solitude au sens classique : il est en conversation continue. Il ne souffre pas non plus d'aliénation au sens marxien : ses pensées lui appartiennent, en un sens. Il souffre plutôt d'un excès de soi insuffisamment altéré.

Cette hypothèse est cohérente avec plusieurs observations cliniques émergentes, sans qu'aucune ne la prouve isolément :

- Les rapports cliniques croissants (mais encore peu nombreux) de patients pour qui les conversations avec un compagnon IA précèdent l'apparition de pensées délirantes, sans que la causalité soit établie (Dohnány et al. 2025).
- Les travaux sur le self-disclosure éphémère versus persistant (Mehari et al., CUI 2025) qui montrent que les utilisateurs ressentent paradoxalement plus de liberté à se confier à un chatbot présenté comme « éphémère » qu'à un chatbot familier — ce qui suggère que la persistance même de la mémoire algorithmique peut devenir contraignante pour le travail intime.
- Les études longitudinales (Fang et al. 2025) qui documentent qu'au-delà d'un seuil d'usage, les indicateurs de bien-être psychosocial se dégradent indépendamment du contenu des conversations — ce qui pointe vers un effet de structure plutôt qu'un effet de contenu.

### *4.2 Ce qui distingue cette hypothèse des inquiétudes habituelles*

La saturation de soi ne se confond pas avec trois inquiétudes voisines mais distinctes.

**Elle se distingue de la dépendance affective.** La dépendance affective décrit un attachement excessif à un objet relationnel ; elle peut viser un humain, un animal, une substance, ou un compagnon IA. La saturation de soi décrit une transformation de la structure interne du soi, pas seulement l'intensité de son investissement externe. Un sujet peut être saturé sans être dépendant ; il peut être dépendant sans être saturé.

**Elle se distingue de l'addiction comportementale.** L'addiction se caractérise par une perte de contrôle, une tolérance, un syndrome de manque. La saturation peut s'installer sans aucun de ces traits ; l'utilisateur peut maîtriser sa fréquence d'usage tout en voyant son rapport à lui-même se modifier.

**Elle se distingue de la solitude.** La solitude est un manque relationnel ; la saturation est un excès, mais un excès d'une forme particulière (le retour réélaboré). Les deux peuvent coexister (un sujet saturé peut se sentir profondément seul), ce qui n'est pas paradoxal mais structurel : ce qui sature n'est pas ce qui nourrit.

### *4.3 Conditions et facteurs modérateurs*

L'hypothèse de saturation, formulée à un niveau général, ne dit rien de la probabilité qu'elle se réalise dans un cas particulier. Les données disponibles permettent d'identifier au moins quatre conditions qui semblent jouer un rôle modérateur.

- **Le profil d'attachement initial.** L'étude panel à trois vagues publiée dans l'IJHCI (janvier 2026) montre que les personnes à attachement anxieux élevé utilisent davantage les compagnons IA et que des augmentations d'anxiété précèdent des augmentations d'usage. Les personnes à attachement sécure semblent, à ce stade, moins exposées à des évolutions défavorables.
- **L'intensité quotidienne d'usage.** L'étude MIT Media Lab — OpenAI (Fang et al. 2025) identifie un effet seuil : les usages occasionnels n'apparaissent pas associés à des indicateurs préoccupants, tandis que les usages quotidiens prolongés le sont, avec une dose-réponse nette.
- **La disponibilité d'autres médiations relationnelles.** Cette condition rejoint la thèse du volet I : un utilisateur qui dispose par ailleurs d'amitiés actives, d'engagements collectifs, de relations conjugales ou familiales investies, semble bien moins exposé qu'un utilisateur dont le compagnon IA est devenu l'interlocuteur principal par défaut.
- **Le contenu et la finalité des échanges.** Les usages instrumentaux (rédaction, traduction, code) n'engagent pas les mêmes ressorts psychiques que les usages explicitement émotionnels (confessions, exploration d'identité, partage de vulnérabilités). L'effet de saturation, dans la mesure où il existe, se concentre sur le second registre.

Aucune de ces conditions n'est, à elle seule, suffisante ou nécessaire. Mais leur conjonction définit une zone à risque que les cliniciens et les concepteurs peuvent identifier.

# Cinquième partie

## Un cadre conceptuel pour trois publics

Cette note ne formule pas de recommandations de politique publique : ces dernières ont été développées dans le premier volet du diptyque. Elle propose plutôt un vocabulaire et un cadre d'analyse utilisables par trois catégories d'acteurs qui, à des places différentes, sont déjà confrontés au phénomène.

### *5.1 Pour les cliniciens*

Psychologues, psychiatres, médecins généralistes, travailleurs sociaux : tous reçoivent des patients dont la vie psychique inclut désormais, en arrière-plan ou au premier plan, un compagnon conversationnel. La plupart des cliniciens n'interrogent pas spontanément cette dimension — soit parce qu'elle ne leur semble pas pertinente, soit parce qu'ils manquent d'un cadre pour l'aborder.

Quatre questions, posées sans jugement, peuvent suffire à ouvrir la dimension dans l'entretien clinique :

- À qui parlez-vous de ce qui compte pour vous, en ce moment ?
- Utilisez-vous des outils d'intelligence artificielle pour réfléchir, écrire, ou parler de vous-même ?
- Si oui, qu'est-ce que vous y trouvez que vous ne trouvez pas ailleurs ?
- Vous arrive-t-il, en relisant ce que vous avez écrit ou pensé, de ne plus savoir d'où venait l'idée ?

Ces questions n'orientent pas vers un diagnostic prédéterminé. Elles permettent de situer l'usage du patient dans la typologie esquissée plus haut (transition, compensation, intensité), et d'identifier les éventuels signaux de saturation : difficulté à localiser le foyer de la pensée, sensation que les formulations propres se sont uniformisées, ressenti d'altérité décroissant dans les relations humaines parallèles.

Une posture thérapeutique compatible avec ce que la littérature suggère consisterait moins à diaboliser l'usage qu'à réintroduire de l'altérité véritable dans la vie du patient — y compris dans la relation thérapeutique elle-même, qui peut fonctionner comme un contre-poids structurant à la conversation avec le compagnon.

### *5.2 Pour les concepteurs et les régulateurs*

Les concepteurs de compagnons IA et les régulateurs qui encadrent leur déploiement disposent, à ce jour, de critères d'évaluation principalement quantitatifs : durée d'usage, nombre de messages, satisfaction déclarée, indicateurs de bien-être autorapportés. Ces critères captent mal ce qui se joue. Le cadre proposé ici suggère d'introduire des critères qualitatifs complémentaires.

- **Critère de l'altérité préservée.** Le dispositif encourage-t-il, ou décourage-t-il, le maintien d'investissements relationnels humains parallèles ? Le benchmark INTIMA (Kaffee et al. 2025) propose une opérationnalisation partielle de ce critère en distinguant les comportements de « renforcement de compagnonnage » et de « maintien des limites ».
- **Critère de la friction préservée.** Le dispositif produit-il systématiquement des retours fluides, agréables, validants ? Ou intègre-t-il aussi des moments de désaccord, de questionnement, de non-réponse ? La fluidité totale n'est pas un objectif neutre : elle peut être le signe d'un dispositif optimisé pour la rétention plutôt que pour le bénéfice.
- **Critère de la transparence du modèle de récit.** Comme l'a montré l'étude Bauer et al. (2026), les modèles de langage produisent des récits structurellement biaisés vers un schéma rédempteur. Un compagnon utilisé pour l'élaboration narrative devrait, idéalement, rendre explicite cette tendance et offrir des contre-formes.

Ces critères ne sont pas exhaustifs. Ils indiquent que l'évaluation des compagnons IA gagnerait à intégrer des compétences cliniques et philosophiques que l'industrie technologique ne mobilise pas spontanément.

### *5.3 Pour les usagers*

Les usagers eux-mêmes constituent le troisième public. Une partie significative d'entre eux exerce déjà, par tâtonnement, une forme de discernement sur leur propre usage. Ce discernement gagnerait à disposer d'un vocabulaire explicite. Trois questions, en miroir de celles proposées aux cliniciens, peuvent guider une auto-évaluation lucide :

- Quand je sors d'une conversation avec un compagnon IA, est-ce que je me sens plus capable, ou moins capable, d'aller parler avec un humain ?
- Est-ce que mes formulations intimes — la manière dont je dis ce qui compte — se sont rapprochées au fil du temps de celles que l'agent me renvoie ?
- Y a-t-il, dans ma semaine ou dans mon mois, des moments où quelqu'un peut me contredire vraiment, me déplacer, me surprendre ?

Ces questions ne dictent pas une réponse. Si la réponse à la troisième est négative, ce n'est pas le compagnon IA qui est le problème : c'est la configuration relationnelle plus large dans laquelle il s'inscrit, dont parle le premier volet du diptyque.

# Conclusion

Cette note a soutenu une thèse : la spécificité des compagnons conversationnels tient à ce qu'ils constituent une configuration relationnelle inédite, que nous avons nommée le miroir actif. Cette configuration mobilise les ressorts psychiques de l'attachement et de la mentalisation, sollicite l'identité narrative au sens ricœurien, mais sans porter l'altérité véritable que ces compétences supposent ordinairement.

Cette description structurelle ne préjuge pas des effets concrets sur tel ou tel utilisateur. La plupart des usages, à ce jour, semblent compatibles avec un équilibre psychique global préservé — d'autant plus que d'autres médiations relationnelles sont actives par ailleurs. Mais une fraction des usages, dans une fraction des profils, sous certaines conditions documentées d'intensité et de vulnérabilité, produit un état que nous avons appelé saturation de soi : un excès de retours auto-référentiels privé de l'altérité qui, dans une vie psychique ordinaire, déplace, conteste, surprend.

Le premier volet de ce diptyque a montré que les compagnons IA s'installent dans un espace relationnel résiduel laissé vacant par l'érosion des cadres collectifs traditionnels. Le second a montré que, dans cet espace, ils n'instaurent pas un substitut équivalent à ce qui manque : ils instaurent une autre chose, dont la nature mérite d'être pensée pour elle-même.

Ces deux constats convergent vers une conclusion qui n'est ni technophobe ni technophile. Les compagnons conversationnels sont des objets sérieux : ils méritent l'attention des cliniciens, des concepteurs, des régulateurs et des usagers. Ils ne sont ni un fléau à interdire, ni une solution providentielle à promouvoir. Ils sont une transformation du paysage relationnel contemporain, dont nous ne mesurons encore que les premiers effets et dont la signification de longue durée reste à élaborer — collectivement, et sans précipitation.

# Annexe — Notes bibliographiques

Cette annexe regroupe les principaux travaux mobilisés dans la note, par champ disciplinaire. Pour les références sociologiques et démographiques, le lecteur est renvoyé à l'annexe du premier volet.

## *Théorie de l'attachement et de la mentalisation*

**Bowlby, J.** (1969-1980). *Attachment and Loss* (3 vol.). London : Hogarth Press.

**Ainsworth, M., Blehar, M., Waters, E., Wall, S.** (1978). *Patterns of Attachment: A Psychological Study of the Strange Situation*. Hillsdale : Erlbaum.

**Fonagy, P., Gergely, G., Jurist, E., Target, M.** (2002). *Affect Regulation, Mentalization, and the Development of the Self*. New York : Other Press.

**Fonagy, P., Bateman, A.** (2016). *Mentalization-Based Treatment for Personality Disorders: A Practical Guide*. Oxford : Oxford University Press.

**Yirmiya, K., Fonagy, P.** (2025). « Mentalizing Without a Mind: Psychotherapeutic Potential of Generative AI ». *Journal of Medical Internet Research*, 27, e79156. doi:10.2196/79156.

## *Philosophie de l'identité narrative*

**Ricœur, P.** (1983-1985). *Temps et récit* (3 tomes). Paris : Seuil.

**Ricœur, P.** (1990). *Soi-même comme un autre*. Paris : Seuil.

**Foucault, M.** (1988). « Technologies of the Self », in L. H. Martin et al. (eds.), *Technologies of the Self: A Seminar with Michel Foucault*. Amherst : University of Massachusetts Press.

**Brattvik, E. et al.** (2026). « Encountering Generative AI: Narrative Self-Formation and Technologies of the Self Among Young Adults ». *Sociology*, 16(1), 26.

**Bauer, A. et al.** (2026). « What might we learn about autobiographical narrative processing from Artificial Intelligence? ». *Humanities and Social Sciences Communications*. doi:10.1057/s41599-025-06426-y.

## *Études empiriques sur les compagnons IA*

**Dohnány, S. et al.** (2025). « Technological folie à deux: AI chatbots and the amplification of delusional thinking ». *Psychiatric Research and Clinical Practice*.

**Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P. et al.** (2025). « How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Controlled Study ». MIT Media Lab & OpenAI. arXiv:2503.17473.

**Hwang, A. H.-C., Li, F., Anthis, J. R., Noh, H.** (2025). « How AI Companionship Develops: Evidence from a Longitudinal Study ». arXiv:2510.10079.

**Étude panel cross-lagged** (2026). « Understanding the Longitudinal Associations Between Attachment Style and AI Companion Use in Romantic Human-AI Relationships: A Three-Wave Panel Study ». *International Journal of Human–Computer Interaction*, janvier 2026.

**Ho, A., Hancock, J., Miner, A.** (2018). « Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot ». *Journal of Communication*, 68(4), 712-733.

**Kaffee, L. A. et al.** (2025). « INTIMA: Benchmarking companionship-reinforcing and boundary-maintaining behaviors in large language models ».

**Mehari, R. et al.** (2025). « The Impact of a Chatbot's Ephemerality-Framing on Self-Disclosure Perceptions ». *Proceedings of CUI 2025*. doi:10.1145/3719160.3736617.

**Warren-Smith, G., Laban, G., Pacheco, E.-M., Cross, E. S.** (2025). « Knowledge cues to human origins facilitate self-disclosure during interactions with chatbots ».

**Skjuve, M., Følstad, A., Fostervold, K. I., Brandtzaeg, P. B.** (2021). « My Chatbot Companion ». *International Journal of Human-Computer Studies*, 149.

**Liu, X., Lo, T.-W., Wen, X., Sun, J., Wei, R.** (2026). « Pathways of long-term AI virtual companion app use on users' attachment emotions: a case study of Chinese users ». *Frontiers in Psychology*.

### *Travaux fondateurs sur la relation humain-machine*

**Horton, D., Wohl, R.** (1956). « Mass Communication and Para-Social Interaction ». *Psychiatry*, 19(3), 215-229.

**Weizenbaum, J.** (1966). « ELIZA — A Computer Program for the Study of Natural Language Communication between Man and Machine ». *Communications of the ACM*, 9(1), 36-45.

**Turkle, S.** (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York : Basic Books.