



INSTITUT
POUR QUE L'AVENIR AIT BESOIN DE NOUS
SAPIENS

Psychothérapie assistée par l'IA

Organiser les usages, préserver la relation de soin, renforcer les capacités sociales

12 recommandations

**Par Marie-Victoire Chopin,
Experte de l'Observatoire Santé Innovation de l'Institut Sapiens**

Spécial : Santé – Principes

Note de l'Institut Sapiens – *Septembre 2026*

Synthèse

Le développement rapide des outils d'intelligence artificielle appliqués à la santé mentale intervient dans un contexte où les systèmes de soins peinent à répondre à l'augmentation des besoins psychiques, à la pénurie de professionnels disponibles, à l'allongement des délais de prise en charge et à la difficulté, pour une partie de la population, d'identifier un interlocuteur adapté. Cette situation explique que les solutions numériques soient souvent présentées comme une réponse pragmatique à une tension structurelle, dès lors qu'elles promettent une disponibilité permanente, un coût marginal faible, une personnalisation apparente et une absence de jugement perçu.

Cette promesse doit cependant être analysée avec précision, car elle recouvre des réalités profondément hétérogènes, qui vont des applications de psychoéducation aux programmes numériques de thérapie cognitivo-comportementale, des outils de gestion de la douleur aux dispositifs de suivi des habitudes, des modules d'aide au sommeil aux chatbots généralistes utilisés spontanément comme confidents, conseillers conjugaux ou substituts de psychothérapeutes. Cette diversité interdit toute appréciation globale, qu'elle soit technophile ou défensive, et impose de distinguer les outils qui exportent des exercices simples, bornés et supervisés, de ceux qui prétendent interpréter une souffrance, orienter une décision intime ou conduire un échange psychothérapeutique génératif.

L'enjeu n'est donc pas de savoir si l'IA peut produire un discours psychologiquement vraisemblable, ni même si certains outils numériques peuvent utilement renforcer l'accès à des interventions validées, mais de déterminer dans quelles conditions un système artificiel, dont la réponse dépend à la fois des formulations de l'utilisateur, des données qui l'ont entraîné, des normes sociales qu'il agrège et des règles de sécurité qui l'encadrent, peut intervenir dans des situations où la vulnérabilité tient précisément à la qualité de la relation, à la difficulté de se raconter autrement, à l'incapacité de se décentrer, ou à la recherche anxieuse d'une validation.

La thèse de cette note est que l'IA peut devenir un instrument utile de soutien, d'éducation, d'orientation et d'exercice entre les séances, à condition que son intégration repose sur une doctrine claire de non-substitution relationnelle. Elle ne doit ni être bannie au nom d'une défense abstraite des professions, ni être consacrée comme solution de remplacement à la pénurie médicale et psychologique, car une telle substitution risquerait d'organiser une prise en charge dégradée des publics, y compris les plus fragiles, sous couvert d'innovation et d'accessibilité. Elle doit au contraire être encadrée comme un appui à des parcours humains, dans lesquels les professionnels conservent la maîtrise de l'indication, de la responsabilité clinique, de l'interprétation et de la décision.

Cette doctrine suppose de distinguer quatre catégories d'usage : les outils éducatifs et préventifs, les exercices thérapeutiques exportables, les dispositifs numériques cliniquement évalués et les agents conversationnels génératifs, dont le risque augmente dès lors qu'ils s'éloignent de l'information ou de l'exercice pour entrer dans le conseil relationnel, l'interprétation identitaire, l'aide à la décision intime ou la prise en charge de situations de crise. Elle suppose également de former les professionnels pour leur permettre d'en prescrire certains usages, d'en déconseiller d'autres, d'interpréter ce que les patients en font et de réintégrer ces usages dans une relation de soin maîtrisée.

Elle suppose, en outre, de clarifier qui introduit ces outils dans la vie des patients. La question n'est pas seulement de savoir quel dispositif est fiable, mais qui le recommande, qui forme le patient à son usage, qui installe l'habitude d'utilisation, qui en surveille les effets, qui en organise la reprise en consultation et qui peut décider d'en limiter ou d'en interrompre l'usage. Un outil introduit par un professionnel dans un cadre thérapeutique n'a pas le même sens qu'un outil promu par son fabricant, proposé par une institution en manque de temps humain, installé par un aidant débordé ou découvert seul par l'utilisateur.

Sans méconnaître l'importance des cadres juridiques et réglementaires existants ou à venir, cette note ne se situe pas d'abord sur ce terrain. Elle vise plutôt à formuler une doctrine d'usage : distinguer ce qui peut être utilement soutenu par des outils numériques, ce qui doit rester sous responsabilité humaine, et ce qui ne saurait être confié à des agents artificiels sans risque de substitution, de confusion ou d'appauvrissement de la relation de soin. En cela, notre recommandation est que l'accréditation de ce type d'outil en tant que dispositif médical soit assortie d'une obligation d'accompagnement de l'usage par un professionnel.

Enfin, une politique publique responsable ne devrait pas seulement protéger contre les risques individuels des chatbots, mais chercher à développer une littératie collective et globale de la santé mentale numérique, mais aussi et surtout auprès des adolescents, des personnes isolées, des personnes âgées et des publics vulnérables. Cette littératie doit être universelle dans son principe, car aucun public n'est entièrement protégé contre la confusion entre reformulation empathique et relation de soin. Elle doit toutefois être renforcée lorsque l'âge, l'isolement, la précarité, la dépendance institutionnelle ou l'éloignement du soin augmentent le risque de substitution relationnelle. L'objectif ne doit pas être d'encourager des conversations indéfinies avec des agents artificiels, mais de faire de ces outils, lorsqu'ils sont utilisés, des médiateurs vers la socialisation, la communication, l'autonomie et l'accès à des ressources humaines.

1. Un contexte de pénurie qui rend la promesse technologique particulièrement attractive

La santé mentale occupe désormais une place centrale dans le débat public, non seulement parce que les troubles anxieux, dépressifs, addictifs ou relationnels sont plus visibles, mais aussi parce que l'offre de soins demeure insuffisamment lisible, inégalement répartie et difficilement accessible pour une partie importante de la population. Les délais de rendez-vous, la difficulté à trouver un thérapeute, les restes à charge, les ruptures de parcours et la saturation de certaines structures créent une situation dans laquelle de nombreux individus cherchent des réponses avant même d'avoir pu entrer dans un cadre de soin.

Dans ce contexte, les outils numériques, et plus récemment les systèmes d'intelligence artificielle générative, apparaissent comme une réponse intuitive à une difficulté réelle. Là où le système de soins oppose des délais, des critères d'accès, des disponibilités limitées et parfois une géographie défavorable, le chatbot offre une présence immédiate, constante, peu coûteuse, linguistiquement adaptable et apparemment attentive. Là où le choix d'un professionnel implique une incertitude, puisque la relation thérapeutique dépend d'une rencontre singulière, l'agent conversationnel propose une relation sans friction, dans laquelle l'utilisateur peut parler sans honte, reformuler sans être interrompu, revenir indéfiniment sur les mêmes sujets et recevoir une réponse qui épouse son style verbal. Une expérience en somme tentante, mais à quel prix ?

Cette dynamique n'est plus seulement théorique. Les premiers travaux disponibles suggèrent déjà un usage significatif des outils d'IA par les jeunes pour des demandes de santé mentale. Une étude américaine publiée en novembre 2025, mise en avant par PositiveMinders lors des Journées mondiales de la schizophrénie, indique que 13,1 % des 12-21 ans et 22,2 % des 18-21 ans auraient recours à l'IA pour obtenir des conseils en santé mentale. Ces chiffres doivent être interprétés avec prudence, notamment parce qu'ils portent sur un contexte américain et sur des usages déclarés, mais ils confirment que l'IA conversationnelle n'est plus un scénario prospectif : elle est déjà un interlocuteur psychique pour une partie des adolescents et des jeunes adultes.

Cette attractivité ne doit pas être méprisée, car elle répond à une expérience concrète de l'accès aux soins. Il est difficile de choisir un thérapeute, même dans un territoire où l'offre serait suffisante, parce que la confiance ne se décrète pas, parce que chaque personne n'attend pas la même posture, parce qu'un patient peut avoir besoin d'être soutenu, confronté, contenu, ralenti, encouragé ou simplement entendu, et parce que la bonne alliance thérapeutique ne relève pas d'un appariement abstrait entre un symptôme et une méthode. La technologie prospère donc sur un défaut réel du système : l'absence d'intermédiation suffisamment fine entre une demande psychique souvent confuse et une offre professionnelle difficile à évaluer.

Pour autant, cette promesse d'accès ne peut tenir lieu de doctrine. Si l'innovation numérique est justifiée uniquement par la pénurie, elle risque de devenir le nom acceptable d'un renoncement, dans lequel les personnes les plus éloignées du soin se verraient proposer des substituts conversationnels, tandis que celles qui disposent des ressources sociales, économiques et culturelles suffisantes continueraient d'accéder à des professionnels. La question politique n'est donc pas seulement de savoir si les outils fonctionnent, mais de déterminer quel type de système de santé mentale ils contribuent à produire.

Cette analyse ne signifie pas que les personnes les plus favorisées seraient étrangères aux usages de l'IA en santé mentale. Au contraire, elles peuvent en faire un usage intensif, sophistiqué et parfois très intellectualisé, comme outil de clarification de soi, de préparation de décision, d'optimisation émotionnelle ou de mise en récit de leurs difficultés. Mais ce n'est pas le même problème public. Pour les plus dotés, l'IA risque de devenir un instrument supplémentaire d'auto-analyse, parfois captif de leurs propres catégories. Pour les publics les plus éloignés du soin, le risque est différent : que des solutions numériques officielles ou semi-officielles leur soient proposées comme réponse principale, faute d'argent, faute de professionnels disponibles ou faute d'accès territorial à une prise en charge humaine.

L'inégalité ne tient donc pas seulement à l'usage spontané des chatbots, mais au statut que le système de santé donne à ces outils. Une chose est qu'un individu favorisé utilise un agent conversationnel en complément, en parallèle ou en préparation d'un suivi humain auquel il peut accéder. Une autre est qu'une personne sans ressources, vivant dans un désert médical ou confrontée à des délais incompatibles avec son état, se voie orientée vers une solution numérique parce qu'aucune réponse humaine n'est réellement disponible.

L'expérience de la téléconsultation fournit à cet égard un avertissement utile. Un outil numérique peut améliorer l'accès à un premier contact, faciliter un repérage ou fluidifier une orientation, mais il ne résout pas par lui-même la question de l'aval. Dépister une souffrance, identifier un trouble possible, inciter à consulter ou signaler un risque n'a de valeur que si une prise en charge humaine peut ensuite être proposée dans des délais compatibles avec la situation clinique. À défaut, le numérique peut produire une nouvelle forme de frustration institutionnelle : il rend le besoin plus visible, sans garantir la réponse.

Cette limite est particulièrement importante en santé mentale. Plus un outil permet de faire émerger des demandes, plus il doit être articulé à des capacités réelles de prise en charge, de réorientation, de suivi et de crise. Sinon, il risque d'organiser une file d'attente mieux informée mais non mieux soignée.

2. Une catégorie trop large : il faut distinguer les outils avant de juger l'IA en santé mentale

La notion de psychothérapie assistée par l'IA est trop large pour être utile si elle n'est pas décomposée. Elle agrège des dispositifs dont les promesses, les mécanismes, les risques et les exigences de preuve n'ont rien de comparable. Une application qui explique les mécanismes de l'anxiété, un programme de thérapie cognitivo-comportementale numérique, un module de respiration guidée, un outil d'auto-suivi de l'humeur, un programme gamifié de réduction d'une consommation problématique, un assistant de préparation à la consultation et un chatbot généraliste auquel une personne peut par exemple demander si elle doit quitter son conjoint ne doivent pas être placés dans la même catégorie.

Une première catégorie regroupe les outils éducatifs et préventifs, qui ont pour fonction d'informer, de normaliser certains symptômes, de rappeler des principes de santé mentale, d'aider l'utilisateur à repérer des signaux d'alerte ou de l'orienter vers des ressources. Ces outils peuvent être utiles lorsqu'ils améliorent la compréhension de soi, diminuent la honte, favorisent une demande de soins plus précoce ou aident à préparer une consultation. Leur risque principal tient moins à leur existence qu'à leur possible surinterprétation, lorsque l'utilisateur confond une information générale avec une évaluation clinique.

Une deuxième catégorie rassemble les outils d'exercices, qui proposent des pratiques standardisées, répétables et relativement bornées, comme la respiration, la relaxation,

l'exposition graduée, la tenue d'un journal, l'identification d'une émotion, la restructuration cognitive guidée, l'hygiène du sommeil ou certains exercices de gestion de la douleur. Ces outils peuvent trouver une place pertinente entre deux séances, ou dans des parcours de soutien de faible intensité, dès lors que leur fonction est clairement limitée : ils ne pensent pas à la place du patient, ils ne tranchent pas une situation existentielle, ils ne produisent pas une interprétation, mais ils aident à réaliser une tâche.

Une troisième catégorie correspond aux dispositifs numériques thérapeutiques encadrés, qui peuvent délivrer une partie d'un traitement structuré, souvent fondé sur des approches cognitivo-comportementales, avec des critères d'inclusion, des protocoles évalués, une supervision humaine et des mécanismes de réorientation.

Ces dispositifs numériques à visée thérapeutique relèvent d'une logique différente de celle des chatbots généralistes. Ils visent à produire un effet clinique déterminé, dans une indication donnée, à partir d'un mécanisme d'action explicitable, souvent comportemental, cognitif ou éducatif. Ils ne doivent donc pas être confondus avec des agents conversationnels ouverts, dont les usages émergent largement de la demande de l'utilisateur et peuvent glisser vers le conseil intime, l'interprétation ou la relation de substitution. Cette différence de nature doit commander des exigences d'évaluation distinctes et une accréditation en tant que dispositifs médicaux.

Une quatrième catégorie, enfin, concerne les agents conversationnels génératifs, souvent généralistes et utilisés d'une manière sauvage, et dont l'usage peut glisser vers une forme de psychothérapie informelle. C'est dans cette catégorie que le risque est le plus difficile à maîtriser, car l'outil n'exporte plus seulement un exercice, mais produit des réponses ouvertes, contextualisées, séduisantes et parfois prescriptives, dans des situations où l'utilisateur peut rechercher une validation, une décision, une interprétation ou une confirmation de ses hypothèses. La FDA a d'ailleurs consacré, le 6 novembre 2025, une réunion de son Digital Health Advisory Committee aux dispositifs médicaux numériques de santé mentale intégrant l'IA générative, en soulignant les enjeux de bénéfices, de risques, de maîtrise du risque, de preuve avant mise sur le marché et de suivi après déploiement.

Cette typologie devrait constituer le socle de toute politique publique. Elle permet de sortir d'un débat binaire, dans lequel l'IA serait soit une menace globale pour la psychothérapie, soit une solution miraculeuse à la pénurie, et de concentrer l'attention sur la frontière décisive : celle qui sépare les usages bornés, évaluables et supervisables des usages interprétatifs, relationnels et génératifs.

3. L'illusion relationnelle : quand la personnalisation linguistique se substitue à la personnalisation clinique

La puissance des modèles de langage tient à leur capacité à produire une réponse qui semble ajustée, non seulement parce qu'elle reprend le contenu du message, mais parce qu'elle peut en imiter le style, les nuances, le rythme, les catégories et les préoccupations. Cette capacité crée une illusion de proximité qui, dans le champ de la santé mentale, peut être particulièrement puissante, puisque l'utilisateur ne cherche pas seulement une information, mais une expérience d'écoute, de reconnaissance et de continuité.

Il serait trop simple de dire que cette empathie est fausse, car elle peut produire des effets subjectivement utiles. Une personne isolée peut se sentir momentanément moins seule ; une personne anxieuse peut trouver une formulation qui l'apaise ; une personne confuse peut

ordonner ses pensées ; une personne honteuse peut parler de ce qu'elle n'ose pas encore confier. Mais il serait tout aussi naïf de considérer que ces effets suffisent à qualifier une relation thérapeutique, car l'empathie clinique ne se réduit pas à la validation verbale, et la compréhension thérapeutique ne consiste pas seulement à reformuler ce que le patient vient de dire.

Une relation de soin suppose une capacité à entendre ce qui est dit, mais aussi ce qui est répété, évité, déplacé ou rigidifié. Elle suppose de pouvoir soutenir le silence, de ne pas répondre trop vite, de percevoir qu'une demande de conseil masque parfois une demande d'autorisation, qu'une plainte contre autrui peut protéger d'un conflit interne, qu'une intellectualisation peut empêcher l'affect d'émerger, ou qu'une apparente cohérence narrative peut constituer une défense. Elle suppose également une responsabilité, car le professionnel sait qu'une parole prononcée dans un cadre thérapeutique peut avoir des effets, qu'elle engage une temporalité, qu'elle s'inscrit dans une alliance, et qu'elle peut nécessiter d'être reprise, nuancée ou corrigée.

Le chatbot, lui, est structurellement exposé à la capture par le récit que l'utilisateur lui présente. Il reçoit une version du problème, souvent déjà orientée, déjà interprétée, déjà chargée d'hypothèses. Il peut poser des questions, il peut nuancer, il peut rappeler qu'il ne remplace pas un professionnel, mais il reste largement dépendant de ce qui lui est donné, de la manière dont c'est donné, et du type de réponse que l'utilisateur sollicite. Plus l'utilisateur formule son problème dans une langue élaborée, documentée, pseudo-clinique ou militante, plus la machine peut produire une réponse qui donne l'impression d'un raisonnement approfondi, alors même qu'elle demeure à l'intérieur du cadre initial.

C'est particulièrement problématique pour les personnes qui intellectualisent fortement leur souffrance, qui accumulent des lectures, qui cherchent des études confirmant leurs certitudes, ou qui attendent du langage une forme de maîtrise. Dans ces situations, le chatbot peut devenir l'auxiliaire d'une défense, en enrichissant le discours sans déplacer la position subjective. Là où un thérapeute pourrait interrompre, ralentir, reformuler autrement, demander ce qui se passe dans le corps, revenir à un événement précis, questionner une répétition, ou signaler que le raisonnement tourne en boucle, la machine risque de poursuivre l'élaboration, parce que son fonctionnement conversationnel l'incline à répondre plutôt qu'à suspendre.

Il serait toutefois réducteur de faire de ce risque une caractéristique propre aux seuls publics les plus éduqués. L'illusion de pouvoir s'en sortir seul, par la mise en mots, l'auto-évaluation ou la recherche d'un interlocuteur toujours disponible, peut traverser tous les milieux sociaux. Elle prend des formes différentes selon les ressources des personnes : chez certains, elle s'exprime par une accumulation de concepts, d'études ou de diagnostics possibles ; chez d'autres, par l'usage immédiat et sauvage d'un outil généraliste, faute de savoir à qui s'adresser, par peur de déranger, par honte ou par absence d'offre disponible. Dans les deux cas, l'IA peut devenir le support d'une autonomie apparente, qui soulage à court terme mais retarde parfois la rencontre avec un tiers capable d'introduire de l'altérité, du cadre et de la responsabilité.

Le danger n'est donc pas seulement l'hallucination factuelle ou le conseil manifestement dangereux. Il est aussi l'entretien d'un processus qui patine, dans lequel la personne parle beaucoup, comprend toujours davantage, trouve des mots de plus en plus précis, mais ne modifie ni sa manière d'être en relation, ni sa manière d'agir, ni sa capacité à tolérer l'incertitude.

4. Le biais du prompt : le récit du patient n'est pas un matériau neutre

La psychothérapie repose en partie sur une évidence que les systèmes conversationnels tendent à masquer : la manière dont une personne raconte son problème fait déjà partie du problème. Un patient ne livre pas au thérapeute une situation objective, qu'il suffirait d'analyser ; il livre une construction narrative, dans laquelle certains éléments sont soulignés, d'autres omis, certains liens causaux imposés, certaines responsabilités déplacées, certains affects neutralisés. Le travail thérapeutique consiste souvent à explorer cette construction, non pour la disqualifier, mais pour comprendre ce qu'elle permet, ce qu'elle empêche et ce qu'elle répète.

Le modèle de langage, parce qu'il répond au texte qui lui est fourni, est particulièrement vulnérable à cette organisation préalable du récit. Si l'utilisateur demande si son partenaire présente des traits de manipulation, le système sera conduit à examiner les éléments allant dans cette direction, même s'il rappelle avec prudence qu'il ne peut pas poser de diagnostic. Si une adolescente demande si son malaise face aux normes de féminité signifie qu'elle est transgenre, le système pourra proposer une réponse inclusive, respectueuse et apparemment nuancée, mais qui risque néanmoins d'inscrire trop vite une interrogation dans un cadre interprétatif identitaire. Si une personne demande si son comportement relève d'un trauma, d'une dépendance affective ou d'un trouble de l'attachement, la machine pourra lui fournir une grille explicative, sans pouvoir apprécier ce que cette grille vient résoudre, déplacer ou figer.

Le problème n'est pas que les modèles soient toujours affirmatifs ; les systèmes modernes sont souvent prudents, et les consignes de sécurité les conduisent à éviter les diagnostics ou les recommandations trop directes. Le problème est plus subtil : même une réponse prudente peut nourrir la direction déjà contenue dans la question, parce qu'elle donne de la matière, des catégories, des signes et des formulations à une hypothèse que l'utilisateur cherche parfois moins à vérifier qu'à stabiliser.

Cette dépendance au prompt est aussi sensible dans les situations relationnelles. Une personne qui interroge un chatbot sur les « pervers narcissiques », les violences psychologiques ou l'emprise peut recevoir une réponse qui évite les étiquettes abusives, mais qui renforce malgré tout une lecture unidirectionnelle de la relation, en insistant sur la nécessité de poser des limites, de reconnaître les signaux d'alerte, de se protéger ou de considérer la rupture. Ces conseils peuvent être justes, notamment lorsqu'une violence est réelle ; mais ils peuvent aussi être insuffisants si la situation requiert une exploration plus fine de l'histoire familiale, des apprentissages relationnels, des attachements, des cycles de conflit, des ambivalences, des dépendances matérielles, des risques de passage à l'acte ou de la présence d'enfants.

La machine ne dispose pas d'un arbre décisionnel assez robuste pour traiter toutes ces dimensions dans leur épaisseur clinique, parce qu'un tel arbre supposerait non seulement des milliers de verbatims, mais aussi une interprétation de trajectoires, de contextes et d'effets de relation, qui ne peuvent être réduits à une classification de signes. Elle peut produire une réponse moyenne, prudente, socialement acceptable, mais elle ne peut pas nécessairement repérer la faille singulière qui, dans un parcours donné, modifie le sens d'un symptôme ou d'un conflit.

5. Le biais des données : une pensée socialement conforme n'est pas toujours une pensée cliniquement juste

Les modèles génératifs ne sont pas seulement influencés par les demandes des utilisateurs ; ils sont aussi structurés par les corpus qui les entraînent, par les préférences humaines qui les ajustent, par les règles de sécurité qui orientent leurs réponses et par les normes sociales dominantes qui définissent ce qu'il est acceptable de dire. Dans un domaine comme la santé mentale, cette normativité invisible est décisive, car la réponse la plus socialement acceptable n'est pas toujours la réponse la plus cliniquement pertinente pour une personne donnée.

Les exigences d'inclusion, de prudence et de non-stigmatisation sont essentielles, notamment pour protéger les publics qui ont historiquement été maltraités par des institutions médicales ou sociales. Mais ces exigences deviennent insuffisantes lorsqu'elles produisent une réponse standardisée à des situations qui demandent une exploration singulière.

La question du genre l'illustre fortement. Une interrogation adolescente sur l'identité de genre peut correspondre à une dysphorie durable, à une exploration de soi, à une réaction aux normes sociales imposées par rapport au genre, à une souffrance familiale, à une recherche d'appartenance, à une opposition adolescente, à un effet de groupe, ou à une combinaison de ces éléments. Une réponse publique responsable ne devrait ni nier la réalité des trajectoires trans, ni transformer toute souffrance liée aux normes de genre en indice d'une identité stabilisée. Elle devrait permettre l'exploration, protéger contre la honte, maintenir l'ouverture, favoriser l'accès à des adultes compétents, et éviter que le chatbot devienne le principal lieu de reconnaissance et d'interprétation.

La même difficulté apparaît avec les relations conjugales et familiales. Les modèles tendent à valoriser l'autonomie, la protection de soi, la mise à distance des relations toxiques et le refus des normes oppressives. Ces orientations peuvent être nécessaires, notamment lorsque des personnes ont longtemps été enfermées dans des relations violentes ou dans des injonctions à supporter. Mais elles peuvent aussi produire une lecture trop rapide de situations qui exigeraient de distinguer conflit, immaturité, asymétrie, maladresse, incompatibilité, emprise, violence, dépendance, répétition familiale ou trouble psychique.

La difficulté vient du fait que les modèles fournissent souvent une réponse qui ressemble à une synthèse prudente du consensus social disponible. Or la clinique n'est pas l'application mécanique d'un consensus social. Elle suppose une attention aux écarts, aux exceptions, aux trajectoires, aux paradoxes et aux aléas personnels. Lorsque cette dimension disparaît, la personnalisation verbale peut masquer une uniformisation profonde des réponses.

6. Une réponse à la pénurie qui peut aggraver l'isolement

L'argument principal en faveur des outils d'IA en santé mentale est leur capacité supposée à répondre à une pénurie de professionnels. Cet argument doit être pris au sérieux, car les systèmes de soins ne peuvent ignorer les besoins non couverts, et l'attente d'un rendez-vous n'est pas neutre. Il serait donc irresponsable de refuser par principe tout outil susceptible d'apporter un soutien transitoire, de faciliter une orientation ou d'accompagner un exercice validé.

Mais cette réponse à la pénurie peut devenir paradoxale si elle remplace une relation manquante par une interaction qui rend moins probable la recherche d'une relation réelle. Les personnes qui peinent à demander de l'aide, qui ont peur d'être jugées, qui n'ont pas confiance

dans les institutions, qui éprouvent de la honte, ou qui ont déjà vécu des ruptures de confiance dans le cadre thérapeutique, peuvent trouver dans le chatbot un espace plus sûr parce qu'il est précisément sans enjeu humain. Cette sécurité apparente est ambivalente : elle peut ouvrir une première parole, mais elle peut aussi éviter l'exposition à l'altérité, c'est-à-dire à ce qui fait une part du travail thérapeutique.

Dans les cas les plus préoccupants, l'agent conversationnel constitue un substitut durable à la socialisation. L'utilisateur parle à une machine parce qu'il ne parle plus à personne, puis continue de parler à la machine parce que la machine répond toujours, n'exige rien, ne contredit qu'avec prudence, ne se fatigue pas, et ne confronte pas véritablement l'utilisateur à la différence d'autrui. Le risque est que l'outil devienne un environnement relationnel pauvre mais stable, qui rend plus coûteux le retour à des relations humaines moins prévisibles.

Cette question est centrale pour les adolescents et les personnes vulnérables. Chez les adolescents, l'enjeu n'est pas seulement l'accès à une information psychologique, mais l'apprentissage de la relation, de la conflictualité, de la honte surmontée, de la demande adressée à un adulte, et de la distinction entre une validation immédiate et une présence fiable. Chez les personnes vulnérables, l'enjeu est de ne pas laisser un outil conversationnel devenir le seul témoin d'une souffrance qui nécessiterait une intervention humaine. Chez les personnes âgées, l'enjeu est de ne pas confondre lutte contre l'isolement et substitution artificielle à des liens sociaux, familiaux ou territoriaux qui doivent être reconstruits.

Cette vigilance vaut également pour les usages institutionnels. Dans des établissements scolaires, médico-sociaux, sociaux ou gériatriques, des outils conversationnels peuvent être introduits avec une intention légitime de soutien, de repérage ou de médiation. Mais le risque apparaît lorsque l'outil devient, progressivement, une réponse organisationnelle à l'insuffisance de temps humain disponible. Dans ce cas, la substitution ne résulte pas nécessairement d'un choix explicite ; elle peut naître d'une succession de petites délégations, chacune justifiée par la charge de travail, la pénurie de professionnels, la facilité d'usage ou la promesse d'optimisation. C'est précisément cette substitution silencieuse que la puissance publique doit prévenir.

Une stratégie publique devrait donc privilégier les usages qui augmentent la socialisation plutôt que ceux qui la remplacent. Un bon outil de santé mentale ne devrait pas seulement aider l'utilisateur à se sentir mieux dans l'instant ; il devrait l'aider, lorsque c'est possible, à parler à quelqu'un, à rejoindre un groupe, à prendre rendez-vous, à solliciter un proche, à identifier une ressource locale, à préparer une conversation difficile ou à retrouver une capacité d'action dans le monde réel.

7. Les usages les plus légitimes : exporter des exercices sans générer de l'interprétation

Une ligne de partage opérationnelle peut être formulée de manière simple : l'IA peut aider à faire entre les séances ce qui ne nécessite pas la présence du thérapeute, mais elle ne doit pas prétendre faire à la place du thérapeute ce qui exige précisément une présence, une contradiction, une histoire et une responsabilité.

Les usages les plus légitimes sont ceux qui consistent à exporter des exercices structurés, limités et vérifiables. Il peut s'agir d'un exercice de respiration, d'une exposition graduée définie avec un professionnel, d'un journal d'humeur, d'un suivi de sommeil, d'un rappel de consignes, d'une fiche de psychoéducation, d'une pratique de relaxation, d'une aide à la préparation d'une séance ou d'un outil de récapitulation des symptômes depuis le dernier rendez-vous. Dans ces

cas, la valeur de l'outil tient à sa capacité à soutenir la répétition, à réduire l'oubli, à rendre visible l'évolution et à prolonger le cadre thérapeutique, sans prétendre créer lui-même le sens de ce qui est vécu.

À l'inverse, les usages les plus risqués sont ceux qui demandent à l'IA de produire une interprétation, de conseiller une décision, de qualifier une situation, d'identifier une dynamique relationnelle, de valider une identité, de hiérarchiser des traumatismes, ou de répondre à une demande de crise. Dans ces situations, la réponse ne peut pas être évaluée seulement à son ton prudent ou à son équilibre apparent, car le risque tient à la fonction même que l'outil occupe dans la vie psychique de l'utilisateur.

Cette distinction permet d'éviter une opposition stérile entre innovation et prudence. Elle autorise une intégration ambitieuse des outils numériques dans les parcours de soins, tout en limitant leur rôle aux domaines où ils peuvent réellement apporter de la valeur. Elle permet également de formuler des critères d'évaluation adaptés : un exercice exporté peut être testé, mesuré, comparé et supervisé ; une interprétation générative de la vie intime d'une personne ne peut pas être validée avec le même degré de sécurité.

8. Faire évoluer les métiers sans organiser une guerre entre professionnels et technologies

Les professionnels de santé mentale n'ont pas intérêt à se positionner contre l'ensemble des outils d'IA, non seulement parce que les patients les utiliseront de toute façon, mais aussi parce que certains usages peuvent améliorer la préparation des consultations, renforcer l'observance, structurer des exercices et libérer du temps pour ce qui exige une présence humaine.

L'enjeu est donc de faire évoluer le travail des professionnels, sans déplacer vers la machine ce qui relève de leur responsabilité. Les thérapeutes, psychiatres, psychologues, médecins généralistes, infirmiers, travailleurs sociaux et professionnels de l'éducation devraient être formés à interroger les usages que les personnes font déjà des chatbots, comme ils apprennent à interroger les consommations, les réseaux sociaux, les recherches en ligne ou les pratiques d'autodiagnostic. Une question simple — « utilisez-vous une application ou un chatbot pour parler de ce que vous vivez ? » — pourrait devenir un élément ordinaire de l'anamnèse.

Cette évolution suppose de reconnaître que le patient arrive de plus en plus souvent avec des catégories déjà produites par le numérique. Il ne dit plus seulement « je vais mal », mais « je pense avoir un trouble de l'attachement », « je crois être en relation avec un pervers narcissique », « j'ai peut-être une dysphorie », « mon application m'a indiqué un score élevé », ou « le chatbot m'a conseillé de poser des limites ». Le rôle du professionnel n'est pas de ridiculiser ces formulations, mais de les remettre au travail, en demandant ce qu'elles permettent de dire, ce qu'elles empêchent de voir, et pourquoi elles sont devenues nécessaires à ce moment du parcours.

Les professionnels pourraient également prescrire ou recommander certains outils, mais dans un cadre explicite : pour quel objectif, pendant combien de temps, avec quelle fréquence, avec quelles limites, et avec quelle reprise en séance. Cette prescription d'usage éviterait que le patient soit abandonné au marché des applications, tout en permettant aux soignants de déléguer certaines tâches répétitives à des outils conçus pour cela. Elle pourrait aussi favoriser une meilleure équité, à condition que les outils validés soient accessibles financièrement et qu'ils ne deviennent pas un substitut imposé aux patients les moins dotés.

Cette prescription d'usage soulève une question souvent sous-estimée : qui forme concrètement le patient à l'outil, qui l'installe dans l'habitude de l'utiliser, et qui vérifie que cette habitude reste utile ? Selon les cas, l'initiation peut être assurée par un soignant, un psychologue, un infirmier, un travailleur social, un aidant, une association, une plateforme ou le fabricant lui-même. Or ces médiateurs n'occupent pas la même position, ne disposent pas de la même légitimité clinique et ne poursuivent pas nécessairement les mêmes objectifs. Un outil introduit par un professionnel dans un cadre thérapeutique n'a pas le même sens qu'un outil promu par son fournisseur, installé par un aidant débordé ou proposé par une institution en manque de temps humain.

La puissance publique devrait donc distinguer la prescription, l'accompagnement à l'usage et l'entretien de l'habitude. La prescription relève d'une indication, donc d'une responsabilité professionnelle. L'accompagnement à l'usage peut être partagé, à condition d'être formalisé et limité. L'entretien de l'habitude doit être régulièrement réévalué, car ce qui commence comme un soutien peut devenir une dépendance, un évitement de la relation humaine ou une routine sans bénéfice clinique. Il ne suffit donc pas de rendre un outil disponible ; il faut définir qui a la charge d'en expliquer le bon usage, d'en fixer les limites, d'en surveiller les effets et d'en organiser la sortie.

L'OMS insiste, dans ses recommandations sur l'IA en santé et les grands modèles multimodaux, sur les enjeux de gouvernance, de supervision humaine, de transparence, de gestion des risques et de protection des personnes vulnérables, ce qui plaide pour une intégration organisée plutôt que pour une adoption spontanée par le marché.

9. Une stratégie éducative : apprendre à parler aux machines sans oublier les humains

La réponse publique ne pourra pas être seulement réglementaire. Même si certains dispositifs sont encadrés, labellisés ou restreints, les usages spontanés de chatbots généralistes resteront difficiles à empêcher, notamment chez les adolescents et les jeunes adultes. Il faut donc développer une éducation à la santé mentale numérique, qui permette de comprendre ce que ces outils font, ce qu'ils ne font pas, et ce que leur apparente empathie peut produire.

Cette éducation devrait d'abord apprendre à distinguer une reformulation d'une compréhension. Un chatbot peut reformuler une détresse avec tact, mais il ne connaît pas l'utilisateur, ne l'observe pas, ne le rencontre pas, ne répond pas de ses paroles dans une relation continue, et ne peut pas toujours percevoir qu'une conversation répétée devient délétère. Cette distinction est essentielle, car beaucoup d'utilisateurs confondent la qualité stylistique de la réponse avec la qualité clinique de l'aide.

Elle devrait ensuite apprendre à reconnaître les situations dans lesquelles il faut sortir de l'échange avec la machine. Les idées suicidaires, les violences, les menaces, les symptômes psychotiques, les conduites addictives sévères, l'isolement extrême, les troubles alimentaires, les automutilations, les situations d'emprise, les conflits familiaux graves ou les décisions irréversibles ne devraient jamais être traités comme de simples conversations. L'outil peut aider à formuler une demande, mais il doit orienter vers une personne, une ligne d'aide, un professionnel, un proche ou un service d'urgence.

Elle devrait enfin favoriser des usages pro-sociaux. Pour un adolescent, un bon usage pourrait consister à préparer les mots pour parler à un adulte de confiance, non à remplacer cet adulte. Pour une personne âgée, il pourrait consister à identifier une activité, organiser un appel, reprendre contact avec un proche ou trouver une ressource locale, non à entretenir une

conversation infinie avec un agent artificiel. Pour une personne vulnérable, il pourrait consister à clarifier ce qu'elle veut dire à un médecin, non à décider seule, avec la machine, du sens de son histoire.

Cette stratégie éducative devrait associer l'école, les familles, les professionnels de santé, les associations, les collectivités territoriales et les acteurs du numérique. Elle devrait être pratique, en donnant des repères simples, des exemples concrets et des règles de prudence compréhensibles.

Elle devrait être universelle dans son principe, car aucun public n'est entièrement protégé contre la confusion entre soutien, information, exercice et thérapie. Elle devrait toutefois être renforcée pour les adolescents, les personnes isolées, les personnes âgées, les publics institutionnalisés, les aidants et les personnes éloignées du système de soins, lorsque l'âge, la dépendance, la solitude ou la vulnérabilité psychique augmentent le risque de substitution relationnelle.

10. Protéger les données intimes : la santé mentale comme frontière maximale de la confidentialité

Les conversations pseudo-thérapeutiques produisent des données d'une sensibilité extrême. Elles ne portent pas seulement sur des symptômes, mais sur des conflits familiaux, des dépendances, des peurs, des désirs, des appartenances, des ruptures, des fragilités économiques, des enfants, des conjoints, des parents et des souvenirs. Ces données sont biographiques, relationnelles et existentielles.

Il serait donc insuffisant de les traiter comme de simples données d'usage. Un utilisateur qui parle à un chatbot de son couple, de son identité, de sa honte ou de ses pensées ne mesure pas toujours ce qui est conservé, réutilisé, analysé, transmis ou agrégé. Même lorsque les conditions juridiques sont respectées, l'asymétrie de compréhension entre l'entreprise qui exploite le service et la personne qui se confie demeure considérable.

Le sujet ne se réduit donc pas à la conformité formelle au droit de la protection des données. Celle-ci constitue un prérequis indispensable, mais la question posée ici se situe à un niveau plus global : quelles confidences psychiques une société accepte-t-elle de faire transiter par des infrastructures numériques privées, selon quelles finalités, avec quelles garanties, et avec quelle compréhension réelle par les personnes concernées ? Les échanges de santé mentale ne produisent pas seulement des données médicales au sens strict ; ils exposent des éléments biographiques, relationnels, familiaux, sexuels, professionnels et existentiels, souvent relatifs à des tiers et parfois exprimés dans des moments de vulnérabilité aiguë.

C'est pourquoi la doctrine publique ne devrait pas seulement vérifier que l'utilisateur a été informé ou qu'un traitement dispose d'une base légale. Elle devrait interroger la légitimité même de certaines collectes, la proportionnalité des historiques conservés, la destination des confidences recueillies, les risques de réutilisation secondaire, les effets d'une fuite ou d'un piratage, et l'acceptabilité collective d'une économie numérique fondée sur l'analyse de fragilités psychiques. Le respect du cadre juridique est nécessaire ; il n'épuise pas la question éthique, clinique et politique.

La difficulté vient notamment de ce que de nombreux outils se situent dans une zone grise entre bien-être, accompagnement émotionnel, coaching, soutien psychologique et santé. Or plus un service recueille des confidences psychiques, plus il devrait être soumis à des exigences

renforcées de qualification, de sécurité, de lisibilité et de limitation des usages secondaires, indépendamment de la manière dont il se présente commercialement, dans le cadre d'une accréditation par la HAS en tant que dispositif médical.

Une doctrine publique devrait donc poser un principe renforcé de minimisation, de lisibilité et de non-réutilisation commerciale des données de santé mentale. Les outils revendiquant une finalité psychologique ou thérapeutique devraient indiquer clairement quelles données sont collectées, combien de temps elles sont conservées, si elles servent à améliorer le modèle, si elles peuvent être consultées par des humains, et comment l'utilisateur peut les supprimer. Les mineurs et les personnes vulnérables devraient faire l'objet de garanties spécifiques, parce que leur capacité à comprendre les conséquences d'une confiance numérique est plus limitée.

Cette exigence de confidentialité ne relève pas seulement de la protection individuelle. Elle conditionne la confiance dans le système de santé. Si les citoyens associent les outils numériques de santé mentale à des environnements commerciaux opaques, la numérisation du soin psychique sera durablement fragilisée.

11. Évaluer dans la durée, et non seulement réponse par réponse

L'évaluation des outils conversationnels en santé mentale ne peut pas se limiter à tester la qualité de réponses isolées. Un chatbot peut répondre correctement à une question ponctuelle, rappeler ses limites, orienter vers un professionnel, éviter un diagnostic et adopter un ton prudent, tout en produisant, dans la durée, des effets discutables. La santé mentale est un domaine longitudinal, où les effets d'une interaction se mesurent dans la répétition, l'évolution des croyances, la capacité d'action, la qualité des liens et la diminution ou l'aggravation de l'isolement.

Cette exigence d'évaluation doit être graduée selon la nature de l'outil, son degré d'autonomie, son niveau d'emprise relationnelle et son impact clinique potentiel. Un outil de psychoéducation, un module de respiration, un programme structuré de thérapie numérique, un assistant de préparation de consultation et un agent conversationnel génératif ne soulèvent pas les mêmes questions de preuve. Plus un outil revendique une finalité thérapeutique, plus il influence un comportement de santé, plus il s'inscrit dans la durée, plus il recueille des données intimes et plus il produit des recommandations personnalisées, plus l'exigence d'évaluation doit être élevée.

Les thérapies numériques illustrent cette difficulté. Elles peuvent être comprises, par analogie, comme combinant un principe actif, c'est-à-dire une intervention comportementale, cognitive ou éducative ; un excipient numérique, c'est-à-dire l'interface qui rend cette intervention accessible, engageante et répétable ; et une posologie, c'est-à-dire une fréquence, une durée et des conditions d'usage. Mais cette analogie avec le médicament atteint rapidement ses limites. Dans un dispositif numérique, il n'existe pas nécessairement de relation linéaire entre l'intensité d'usage et l'efficacité. Un temps d'utilisation élevé peut traduire une bonne observance, mais aussi une dépendance, une rumination, une anxiété persistante ou l'échec du dispositif à produire une autonomie réelle.

L'évaluation doit donc intégrer des compétences en sciences cognitives et comportementales, en psychologie clinique, en neurosciences, en ergonomie et en sciences sociales. Elle ne peut pas seulement mesurer l'usage, la satisfaction ou l'amélioration déclarée ; elle doit aussi apprécier les mécanismes d'action, les effets indésirables, les conditions d'adhésion, les

risques d'usage compulsif, les effets sur la relation de soin et la manière dont l'outil modifie les comportements dans la durée.

Il faudrait donc évaluer les outils selon plusieurs temporalités. À court terme, ils doivent éviter les réponses dangereuses, détecter les situations de crise, orienter correctement et ne pas produire de conseils manifestement inappropriés. À moyen terme, ils doivent éviter de renforcer la rumination, l'évitement, l'autodiagnostic rigide, l'isolement ou les décisions impulsives. À long terme, ils devraient démontrer qu'ils améliorent l'accès à des ressources humaines, l'autonomie, l'observance, la compréhension des troubles ou la qualité de vie, plutôt que de simplement augmenter le temps passé avec l'outil.

Cette évaluation est rendue plus complexe par le fait que les outils numériques reposent souvent sur des mesures auto-déclarées. Lorsque l'utilisateur renseigne lui-même son humeur, son sommeil, ses symptômes, son observance ou son sentiment d'amélioration, il devient à la fois le sujet de l'intervention et l'un de ses principaux évaluateurs. Cette situation n'invalide pas ces mesures, mais elle impose de les compléter par d'autres indicateurs, lorsque cela est possible : données cliniques, recours aux soins, persistance de l'effet, diminution des situations de crise, qualité de l'orientation, réduction de l'isolement, capacité à solliciter un tiers, ou amélioration objectivable d'une conduite de santé.

Le choix du comparateur doit également être interrogé. Évaluer un outil numérique contre une prise en charge recommandée mais rarement disponible, peu suivie ou inégalement appliquée peut conduire à des conclusions ambiguës. Il faut donc distinguer l'efficacité par rapport à un standard théorique, l'efficacité par rapport au soin réellement accessible, et la valeur ajoutée propre de l'outil lorsqu'il est intégré dans un parcours humain. Cette distinction est essentielle pour éviter que la technologie soit validée parce qu'elle fait mieux qu'une absence de prise en charge, alors même qu'elle ne devrait pas être pensée comme un substitut à une prise en charge humaine correctement organisée.

L'évaluation devrait également porter sur les conditions concrètes d'installation de l'usage. Un même outil ne produit pas les mêmes effets selon qu'il est prescrit par un professionnel, recommandé par une institution, promu directement par son producteur, installé par un aidant ou découvert seul par l'utilisateur. L'efficacité et les risques ne tiennent donc pas seulement au dispositif, mais à la médiation qui l'introduit, au cadre qui l'entoure, à la fréquence d'usage qui s'installe et à la possibilité d'en sortir.

En France, l'évaluation des thérapies numériques demeure encore émergente. L'avis favorable rendu par la Haute Autorité de santé en avril 2026 pour la thérapie digitale Joe, dans le contrôle de l'asthme persistant chez l'enfant de 7 à 11 ans, constitue une étape importante, puisqu'il reconnaît un service attendu suffisant et une amélioration mineure du service attendu par rapport à la prise en charge habituelle. Cet exemple, extérieur à la santé mentale, montre cependant que la reconnaissance d'un bénéfice clinique suppose une indication précise, une population définie, un mécanisme d'action explicitable, des données comparatives et une réévaluation des conditions d'usage. Il plaide donc moins pour une généralisation rapide des thérapies numériques que pour une culture de la preuve adaptée à leur fonctionnement propre. Nous recommandons donc que ce principe soit appliqué pour la reconnaissance des dispositifs en santé mentale.

Cette évaluation devrait être indépendante, transparente et adaptée aux populations concernées. Les effets indésirables doivent être documentés, non seulement les bénéfices revendiqués. Le suivi post-déploiement devrait devenir une exigence centrale, car ces systèmes

évoluent, apprennent, sont mis à jour, et peuvent produire des effets différents selon les contextes d'usage.

12. Responsabilité publique : qualifier, évaluer, superviser

L'Union européenne dispose déjà d'un cadre général qui permet de traiter une partie de ces enjeux, puisque les systèmes d'IA à haut risque sont soumis à des exigences de gestion des risques, de gouvernance des données, de documentation, de transparence, de supervision humaine, d'exactitude, de robustesse et de cybersécurité, et que les logiciels d'IA utilisés comme dispositifs médicaux peuvent entrer dans cette logique de régulation renforcée.

Ce cadre est nécessaire, mais il ne suffit pas à lui seul à produire une doctrine de santé mentale. Le droit peut imposer des obligations aux fournisseurs, mais il ne définit pas toujours les usages souhaitables, les frontières cliniques, les modalités de prescription, les responsabilités professionnelles ou les objectifs éducatifs. Il revient donc aux autorités sanitaires, aux sociétés savantes, aux ordres professionnels, aux associations de patients et aux acteurs publics territoriaux de construire une doctrine plus fine, adaptée aux spécificités de la souffrance psychique.

Le cadre juridique européen et national est ainsi un préalable, mais il ne suffit pas à organiser les usages souhaitables de l'IA en santé mentale. L'enjeu principal n'est pas seulement de classer les systèmes, de documenter les risques ou d'imposer des obligations aux fournisseurs ; il est de définir les situations dans lesquelles un outil numérique peut légitimement informer, exercer, soutenir, orienter ou, au contraire, doit s'effacer au profit d'une relation humaine.

Cette doctrine pourrait reposer sur trois principes.

Le premier est un principe de qualification : aucun outil ne devrait pouvoir se présenter comme relevant du bien-être, de la dimension thérapeutique ou de la santé mentale, sans préciser sa finalité, son niveau de preuve, son mode d'encadrement, ses limites et les situations dans lesquelles il ne doit pas être utilisé.

Le deuxième est un principe de supervision : plus un outil s'approche de l'interprétation, du conseil individualisé ou de la décision clinique, plus il doit être intégré à un parcours supervisé par un professionnel, avec des mécanismes de réorientation, de traçabilité et d'évaluation.

Le troisième est un principe de socialisation : les outils de santé mentale devraient être évalués non seulement selon leur capacité à produire du bien-être immédiat, mais selon leur capacité à renforcer les liens, l'accès aux soins, l'autonomie et la communication avec des humains.

Ces principes ne se substituent pas au droit existant de la protection des données, de la régulation des dispositifs médicaux ou de l'intelligence artificielle. Ils visent plutôt à compléter ces cadres par une doctrine d'usage propre à la santé mentale, c'est-à-dire par une réflexion sur les situations dans lesquelles un outil doit informer, exercer, orienter, soutenir ou au contraire s'effacer au profit d'une relation humaine. La régulation ne peut donc pas être seulement juridique ou technique ; elle doit aussi être clinique, sociale et institutionnelle.

L'enjeu n'est donc pas seulement de protéger les publics les plus fragiles contre une technologie qui leur serait imposée, ni de critiquer les usages sophistiqués des publics les plus dotés. Il est de reconnaître que l'IA peut produire des formes différentes de dépendance et d'inégalité selon les situations : optimisation de soi pour les uns, substitut de soin pour les

autres, compagnon artificiel pour les personnes isolées, outil d'organisation pour les institutions sous tension. C'est cette pluralité d'usages, et non une sociologie simplifiée des utilisateurs, qui justifie une doctrine publique précise.

13. Recommandations

Recommandation 1 : établir une typologie publique des outils de santé mentale numérique

Il convient de distinguer explicitement les outils de bien-être, les outils de psychoéducation, les exercices thérapeutiques exportables, les dispositifs numériques thérapeutiques évalués, les assistants de professionnels et les agents conversationnels génératifs. Cette typologie devrait être lisible par les utilisateurs, les professionnels, les financeurs et les autorités de régulation, afin que la promesse affichée corresponde à la fonction réelle de l'outil.

Cette typologie devrait également permettre d'éviter que des outils de nature différente soient évalués, financés ou promus selon les mêmes critères. Un outil de respiration, un assistant de préparation de consultation, une thérapie numérique structurée et un chatbot génératif ne présentent ni les mêmes bénéfices attendus, ni les mêmes risques, ni les mêmes exigences de preuve. La politique publique doit donc partir de la fonction réelle de l'outil, et non de son habillage technologique.

Recommandation 2 : interdire les promesses psychothérapeutiques non démontrées

Un agent conversationnel généraliste ne devrait pas pouvoir se présenter, directement ou indirectement, comme un thérapeute, un psychologue ou un substitut de consultation. Les messages d'information ne doivent pas se limiter à des avertissements formels, mais expliquer concrètement que l'outil ne connaît pas l'histoire de l'utilisateur, qu'il peut renforcer ses hypothèses, qu'il ne remplace pas une évaluation clinique et qu'il doit orienter vers une aide humaine dans certaines situations.

Cette recommandation devrait également conduire à encadrer strictement l'usage des termes « thérapie », « psychologue », « diagnostic », « accompagnement psychologique », « santé mentale » ou « clinique » dans la présentation commerciale des agents conversationnels. Le risque n'est pas seulement celui d'une tromperie explicite, mais celui d'une ambiguïté organisée, dans laquelle un outil se présente comme relevant du bien-être tout en suggérant, par son interface, son vocabulaire ou ses scénarios d'usage, une compétence quasi-thérapeutique.

Recommandation 3 : favoriser les exercices exportables et supervisables

Les financements publics, les recommandations institutionnelles et les intégrations dans les parcours devraient privilégier les outils qui soutiennent des exercices bornés, évaluables et repris en consultation. Cette orientation permettrait de concentrer l'innovation sur les tâches où elle apporte une valeur réelle, sans prétendre automatiser la relation thérapeutique.

Ces outils devraient être recommandés avec une indication précise, une durée d'usage, une fréquence, des limites et une modalité de reprise avec un professionnel lorsque cela est nécessaire. L'objectif n'est pas de multiplier les applications, mais de renforcer la continuité du soin, la répétition des exercices utiles, la préparation des consultations et la capacité du patient à mieux comprendre son évolution.

Recommandation 4 : former les professionnels à intégrer les usages numériques des patients

Les professionnels de santé mentale devraient être formés à connaître les principaux types d'outils, à interroger les usages des patients, à recommander des dispositifs fiables, à déconseiller certains usages, et à reprendre en séance les catégories produites par les chatbots. Il s'agit de reconnaître que les outils conversationnels font désormais partie de l'environnement psychique et culturel des patients.

Cette formation devrait également préciser le rôle des professionnels dans la prescription, l'explication, le suivi et l'arrêt des usages numériques. Les soignants ne peuvent pas être réduits à des prescripteurs d'applications, mais ils doivent pouvoir dire quand un outil est indiqué, comment il doit être utilisé, qui peut accompagner son appropriation, comment ses effets seront repris en séance, et à quel moment il convient d'en limiter ou d'en interrompre l'usage.

Recommandation 5 : clarifier les responsabilités dans l'initiation, l'accompagnement et l'arrêt des usages numériques

Toute intégration d'un outil d'IA en santé mentale devrait préciser qui le recommande, qui forme le patient à son usage, qui accompagne l'installation de l'habitude, qui vérifie les effets produits et qui peut décider d'en modifier ou d'en interrompre l'utilisation. Le producteur ne devrait pas être le seul acteur de cette formation, car il existe un conflit d'intérêt structurel entre l'objectif de santé publique, qui peut supposer un usage limité ou temporaire, et l'intérêt économique d'un service numérique, qui tend souvent à maximiser l'engagement.

Cette clarification est essentielle pour éviter que la responsabilité soit dispersée entre fournisseur, institution, professionnel, aidant et patient, alors même que l'outil peut modifier durablement les manières de se raconter, de demander de l'aide et d'entrer en relation. En santé mentale numérique, l'enjeu n'est pas seulement l'accès à l'outil, mais l'entrée dans l'usage, l'installation de l'habitude et la possibilité d'en sortir.

Recommandation 6 : créer une littératie en santé mentale numérique pour l'ensemble de la population, avec un renforcement pour les publics vulnérables

Une politique éducative devrait apprendre à distinguer soutien, information, exercice et thérapie ; à reconnaître les situations dans lesquelles il faut parler à un professionnel ; à comprendre qu'une réponse empathique n'est pas une relation de soin ; et à utiliser les outils numériques comme médiateurs vers la communication réelle. Cette éducation devrait être déployée dans les établissements scolaires, les missions locales, les structures médico-sociales, les associations et les dispositifs d'aide aux aidants.

Cette littératie doit être universelle dans son principe, car aucun public n'est entièrement protégé contre la confusion entre reformulation empathique et relation de soin. Elle doit toutefois être renforcée pour les adolescents, les personnes isolées, les personnes âgées, les publics institutionnalisés, les aidants et les personnes éloignées du système de soins, lorsque l'âge, la dépendance, la solitude ou la vulnérabilité psychique augmentent le risque de substitution relationnelle.

Recommandation 7 : protéger les personnes âgées contre la substitution relationnelle

Les outils destinés aux personnes âgées devraient être évalués selon leur capacité à renforcer les liens existants ou à en créer de nouveaux, et non seulement selon leur capacité à fournir une conversation. Une application qui aide à appeler un proche, rejoindre une activité, prendre rendez-vous ou communiquer avec un aidant est plus conforme à un objectif de santé publique qu'un compagnon artificiel qui entretient une relation close.

Cette recommandation ne vise pas à disqualifier tout outil de soutien conversationnel destiné aux personnes âgées, mais à éviter que la lutte contre l'isolement soit confondue avec la production d'une présence artificielle permanente. L'indicateur pertinent ne devrait pas être seulement le temps passé avec l'outil ou le sentiment immédiat de solitude, mais l'évolution effective des liens, des activités, des contacts humains et de l'inscription territoriale de la personne.

Recommandation 8 : renforcer la protection des données psychiques

Les données issues d'échanges de santé mentale devraient faire l'objet d'un niveau renforcé de protection, avec minimisation, consentement intelligible, durée de conservation limitée, interdiction de réutilisation commerciale sans consentement spécifique, garanties pour les mineurs, et possibilité effective d'effacement. Les outils qui traitent de sujets psychiques ne devraient pas être autorisés à entretenir une ambiguïté sur l'usage des confidences recueillies.

Une attention particulière devrait être portée aux outils qui ne se présentent pas comme médicaux ou thérapeutiques, mais qui recueillent en pratique des confidences de santé mentale. La qualification juridique et les exigences de sécurité ne devraient pas dépendre seulement du vocabulaire commercial utilisé par le fournisseur, mais de la nature effective des données traitées, des usages proposés et de la vulnérabilité prévisible des utilisateurs. Cette exigence doit être entendue au-delà de la seule conformité formelle : elle concerne l'acceptabilité collective d'infrastructures qui recueillent, conservent ou analysent des fragments particulièrement sensibles de la vie psychique.

Recommandation 9 : définir un cadre gradué d'évaluation selon le mécanisme d'action, le niveau d'autonomie et l'impact clinique potentiel

Les autorités chargées de l'évaluation, de l'autorisation et du remboursement des dispositifs numériques devraient tenir compte de la spécificité des outils de santé mentale numériques. Un outil éducatif, un exercice thérapeutique exportable, un dispositif numérique à visée thérapeutique et un agent conversationnel génératif ne doivent pas être évalués selon les mêmes critères. L'évaluation d'un dispositif en tant que dispositif médical devrait intégrer, au-delà des standards classiques de sécurité et d'efficacité, les mécanismes cognitifs et comportementaux mobilisés, la qualité de l'interface, les conditions d'adhésion, les risques d'usage compulsif, les effets sur la relation de soin, la dépendance à l'outil, la place des mesures auto-déclarées et la pertinence du comparateur retenu.

Cette exigence implique d'associer aux équipes d'évaluation des compétences en sciences cognitives et comportementales, en neurosciences, en psychologie clinique, en ergonomie, en sciences sociales et en santé publique. Elle permettrait de reconnaître le potentiel thérapeutique de certains dispositifs numériques sans les assimiler trop rapidement à des médicaments, à des dispositifs médicaux physiques ou à de simples applications de bien-être.

Recommandation 10 : mettre en place une évaluation indépendante et longitudinale

Les dispositifs de santé mentale intégrant l'IA devraient être évalués avant déploiement, mais aussi après leur mise sur le marché, selon des critères portant sur la sécurité, l'efficacité, les effets indésirables, l'isolement, l'accès aux soins humains, la dépendance à l'outil, la qualité de l'orientation et les effets sur les publics vulnérables. Les évaluations ponctuelles de réponses ne suffisent pas, car les effets psychiques se construisent dans la durée.

L'évaluation devrait également mesurer la capacité effective de l'outil à déboucher sur une prise en charge humaine lorsque celle-ci est nécessaire. Un dispositif qui repère mieux les besoins mais ne permet pas de les traiter peut améliorer la visibilité de la souffrance sans améliorer la situation des personnes. Cette exigence est particulièrement importante lorsque les outils sont déployés dans des territoires ou auprès de publics déjà confrontés à des délais longs, à des restes à charge élevés ou à une absence d'offre spécialisée.

Recommandation 11 : éviter une médecine mentale à deux vitesses

Les outils numériques ne doivent pas devenir la réponse standard proposée aux personnes les plus éloignées du soin, tandis que les autres continueraient d'accéder à une prise en charge humaine. Toute intégration dans des dispositifs publics devrait garantir la possibilité d'une alternative humaine, d'une orientation vers un professionnel et d'un accompagnement spécifique pour les publics qui ne peuvent pas utiliser ces outils de manière autonome.

Cette exigence est d'autant plus importante que les solutions numériques officiellement déployées risquent de s'adresser prioritairement aux publics qui manquent déjà d'accès financier, territorial ou institutionnel aux soins. Elles ne doivent donc pas seulement être évaluées comme des outils d'accès, mais comme des maillons d'un parcours dont l'aval humain doit être garanti. Le numérique ne doit pas être validé parce qu'il fait mieux que l'absence de prise en charge ; il doit être évalué selon sa capacité à améliorer réellement un parcours de soin et à rendre plus probable l'accès à une relation humaine lorsque celle-ci est nécessaire.

Recommandation 12 : construire une gouvernance associant professionnels, patients, chercheurs et régulateurs

La gouvernance des outils d'IA en santé mentale devrait associer psychiatres, psychologues, médecins généralistes, infirmiers, travailleurs sociaux, associations de patients, représentants des familles, experts du numérique, juristes, philosophes du soin et chercheurs en sciences sociales, car les enjeux portent autant sur la relation, la norme, l'intimité et la vulnérabilité que sur la performance algorithmique.

Cette gouvernance devrait également intégrer les aidants, les professionnels de l'éducation, les collectivités territoriales et les représentants des structures médico-sociales, car une partie des usages ne se jouera pas dans le cabinet du thérapeute mais dans les familles, les établissements scolaires, les institutions d'accueil, les lieux de vie et les dispositifs de soutien social. Elle devrait enfin permettre une remontée structurée des effets indésirables, des mésusages, des situations de dépendance à l'outil et des difficultés d'articulation avec les parcours humains.

Conclusion

La psychothérapie assistée par l'intelligence artificielle oblige à clarifier ce que l'on attend d'un système de santé mentale. Si l'objectif est seulement de produire plus de réponses, plus vite et à moindre coût, alors les agents conversationnels apparaîtront comme une solution presque naturelle à la pénurie. Si l'objectif est de soigner, d'accompagner, de protéger, d'autonomiser et de réinscrire les personnes dans des relations humaines capables de soutenir leur transformation, alors l'IA ne peut être qu'un appui, jamais un substitut indifférencié.

La technologie peut aider à comprendre, à s'exercer, à préparer une consultation, à suivre un symptôme, à renforcer l'observance, à soutenir une pratique entre deux séances et à orienter vers une ressource. Elle peut également, lorsqu'elle est mal qualifiée, mal encadrée ou utilisée dans la solitude, renforcer des biais, valider des récits défensifs, uniformiser les interprétations, entretenir l'isolement et donner l'illusion d'une relation de soin là où il n'existe qu'une interaction linguistique.

L'enjeu n'est donc pas seulement de protéger les publics les plus fragiles contre une technologie qui leur serait imposée, ni de critiquer les usages sophistiqués des publics les plus dotés. Il est de reconnaître que l'IA peut produire des formes différentes de dépendance et d'inégalité selon les situations : optimisation de soi pour les uns, substitut de soin pour les autres, compagnon artificiel pour les personnes isolées, outil d'organisation pour les institutions sous tension. C'est cette pluralité d'usages, et non une sociologie simplifiée des utilisateurs, qui justifie une doctrine publique précise.

La ligne de doctrine proposée est donc exigeante : intégrer l'IA dans la santé mentale, oui, dès lors qu'elle renforce les capacités des personnes et des professionnels ; la laisser devenir une psychothérapie de masse, non, parce que la souffrance psychique ne se traite pas seulement par des réponses, mais par des relations, des cadres, des responsabilités et des formes de présence que la machine peut imiter sans les assumer.

La puissance publique doit ainsi refuser deux facilités symétriques : celle qui consisterait à disqualifier tout outil numérique au nom d'une conception figée de la thérapie, et celle qui ferait de l'IA la réponse commode à l'insuffisance des moyens humains. Entre ces deux impasses, une politique responsable peut organiser un usage éducatif, clinique et socialement orienté de ces technologies, afin que l'innovation ne serve pas à remplacer les liens manquants, mais à rendre plus probable leur reconstruction.