**Introduction générale à l’intelligence artificielle et à la psychologie clinique**

**Licence 3 TD UE 506**

La révolution numérique a profondément transformé la manière dont nous appréhendons la santé mentale. L’intelligence artificielle (IA) n’est plus un simple concept de science‑fiction ; elle est devenue un ensemble d’outils qui apprennent à partir de données, raisonnent et parfois créent. Pour initier des étudiants en psychologie clinique à ces technologies, il est essentiel de comprendre les différentes catégories d’IA et leurs implications pour la pratique clinique. On distingue généralement trois branches :

* **IA symbolique et logique.** Historiquement, les premiers systèmes experts des années 1970–1980 reposaient sur des règles explicites codées par des spécialistes. Ces programmes étaient performants pour résoudre des problèmes bien définis (diagnostic de troubles hématologiques, par exemple), mais rigides dès qu’ils quittaient le cadre prévu. Ils ont nourri les premières réflexions éthiques sur l’autonomie des machines et la responsabilité professionnelle.
* **Apprentissage automatique supervisé et non supervisé.** Dans le *machine learning* moderne, les algorithmes apprennent des relations à partir d’exemples. L’apprentissage supervisé utilise des données étiquetées pour prédire des classes (diagnostic) ou des valeurs continues (nombre de séances nécessaires). L’apprentissage non supervisé, quant à lui, détecte des structures latentes sans labels préalables ; il peut servir à regrouper des patients en sous‑types symptomatiques. On peut y inclure les méthodes de réduction de dimension (analyse en composantes principales) et les techniques d’apprentissage par renforcement qui optimisent des décisions séquentielles (par exemple, choisir entre divers modules thérapeutiques en fonction de la progression du patient).
* **Apprentissage profond et IA générative.** L’*apprentissage profond* s’appuie sur des réseaux de neurones profonds capables d’extraire automatiquement des représentations complexes à partir de données brutes (images, signaux physiologiques, textes). Les versions les plus médiatisées sont les modèles de langage de grande taille (*LLMs*) comme ChatGPT ou Gemini. À partir de billions de paramètres, ces modèles prédisent le mot suivant dans une phrase et génèrent des textes cohérents. L’IA générative se décline en outils de génération d’images, de musique ou de vidéos. Dans le champ de la psychologie clinique, ces modèles peuvent synthétiser des notes de séance, concevoir des stimuli thérapeutiques ou accompagner les patients dans des exercices d’écriture.

L’IA actuelle est dite *faible* car elle est orientée vers des tâches spécifiques. Parler d’une « IA forte » capable de conscience ou d’empathie relève encore de la spéculation. Toutefois, la multiplication des outils, l’augmentation constante des données disponibles et l’intégration de l’IA dans des dispositifs connectés font émerger des questions inédites : comment préserver la confidentialité des échanges ? Comment éviter les biais ? Comment co‑concevoir des systèmes avec les patients ?

**Un bref panorama historique**

La psychologie et l’IA ont entretenu des rapports constants depuis les débuts de l’informatique cognitive. Les premiers systèmes tels que *ELIZA* (1966) simulaient un psychothérapeute rogerien en reflétant les phrases de l’utilisateur. Bien qu’il s’agisse d’un simple script, l’effet ELIZA a mis en évidence la tendance des utilisateurs à anthropomorphiser les machines. Dans les années 1980, des systèmes experts comme MYCIN (diagnostic d’infections sanguines) ou INTERNIST‑I ont inspiré des tentatives en santé mentale, mais la complexité des troubles et l’ambiguïté des symptômes limitaient leur succès. Ce n’est qu’avec l’essor du *machine learning* et du *deep learning* que des avancées significatives ont été observées.

Aujourd’hui, la psychologie clinique se trouve à un carrefour où l’IA peut analyser de vastes ensembles de données, repérer des signaux subtils et générer des outils interactifs. Pour autant, ces technologies ne remplacent pas le jugement clinique. La relation thérapeutique demeure un pilier insubstituable. Les algorithmes restent des *assistants* qui, correctement utilisés, peuvent optimiser l’accès aux soins, personnaliser les interventions et favoriser la recherche.

**Principes de base du *machine learning***

Pour appréhender les usages futurs, il est utile de décomposer le processus du *machine learning* :

1. **Collecte et préparation des données.** Les données peuvent provenir de questionnaires cliniques, de biosignaux (rythme cardiaque, galvanic skin response), de traceurs numériques (géolocalisation, utilisation du smartphone) ou de transcriptions de séances. La qualité des prédictions dépend directement de la qualité et de la représentativité des données.
2. **Choix de modèle.** On distingue des modèles simples (régression logistique, arbres de décision) et des modèles complexes (forêts aléatoires, réseaux de neurones convolutifs). Les études sur la prédiction du trouble de stress post‑traumatique (TSPT) montrent que des arbres de décision peuvent atteindre des AUC élevés mais que leur généralisation est incertaine si l’échantillon est restreint[bmcmedinformdecismak.biomedcentral.com](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02754-2#:~:text=Abstract).
3. **Entraînement et validation.** La base de données est divisée en sous‑ensembles (ensemble d’entraînement, ensemble de validation, ensemble de test) afin d’évaluer la performance et d’éviter le surapprentissage. Les méta‑analyses soulignent l’importance de validations externes pour limiter le risque de biais[bmcmedinformdecismak.biomedcentral.com](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02754-2#:~:text=and%20the%20remaining%20had%20unclear,world%20settings).
4. **Évaluation et interprétation.** Les indicateurs (AUC, précision, rappel, *F1‑score*) doivent être interprétés dans un contexte clinique. Une AUC élevée ne garantit pas que l’outil sera utile ou éthique. Par ailleurs, des méthodes d’explicabilité (ex. SHAP, LIME) permettent d’identifier les variables qui influencent la décision et de détecter des biais implicites.

**Psychologie clinique et psychotraumatisme : le contexte**

**Définitions et spécificités du traumatisme**

Le traumatisme psychique désigne l’effet durable d’un événement ou d’une série d’événements menaçant l’intégrité physique ou psychique. Le trouble de stress post‑traumatique (TSPT) se caractérise par des symptômes regroupés en quatre clusters : reviviscences intrusives (flashbacks, cauchemars), évitement, altérations négatives de la cognition et de l’humeur (culpabilité, dévalorisation, pensées négatives persistantes) et hyper‑réactivité (surveillance excessive, sursauts, irritabilité). La méta‑analyse de Vali et al. rappelle que le TSPT touche jusqu’à 8 % de la population générale et que le coût économique en termes de perte de productivité et de soins de longue durée est élevé[bmcmedinformdecismak.biomedcentral.com](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02754-2). Les formes complexes (CPTSD) intègrent des symptômes additionnels d’altération de la régulation émotionnelle et de l’image de soi.

Dans la pratique, les cliniciens s’appuient sur des guides fondés sur les preuves (ISTSS, OMS, APA), qui recommandent des thérapies centrées sur le traumatisme, notamment la thérapie de traitement cognitif (CPT) et l’exposition prolongée. Pourtant, l’accès à ces thérapies est inégal et les taux d’abandon restent élevés, notamment en raison des difficultés à tolérer l’exposition et du manque de thérapeutes formés[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11797152/#:~:text=Posttraumatic%20stress%20disorder%20,Recent%20advancements%20in). C’est dans ce contexte que l’IA et les technologies numériques offrent des perspectives pour soutenir le dépistage, la prévention et l’intervention.

**Pourquoi l’IA intéresse-t-elle la psychotraumatologie ?**

Outre l’amélioration de l’accès aux soins, plusieurs facteurs motivent l’exploration des outils d’IA en psychotraumatologie :

* **Volume et complexité des données.** Les traumatismes laissent des traces multiformes : altérations du rythme circadien, dysrégulation neuroendocrinienne, récits subjectifs, comorbidités. Les modèles de ML peuvent intégrer ces sources hétérogènes et identifier des profils de risque ou de résilience. Par exemple, en analysant les données de milliers de vétérans, un modèle peut prédire la probabilité d’évolution vers un TSPT chronique et recommander une intervention précoce.
* **Besoin d’outils de triage et de personnalisation.** Les ressources en santé mentale sont limitées. Le triage automatisé peut orienter les patients vers des professionnels spécialisés ou des programmes en ligne adaptés. À terme, l’IA pourrait contribuer à adapter l’intensité et la séquence des interventions en fonction de la progression de chaque individu.
* **Soutien entre les séances.** De nombreux patients peinent à réaliser les devoirs de thérapie, par exemple écrire leurs pensées traumatiques ou pratiquer l’exposition in vivo. Les assistants conversationnels peuvent fournir des rappels, proposer des reformulations et encourager la progression[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11797152/#:~:text=Shinkfield%2C%C2%A02007%20%3B%20Reger%20et%C2%A0al,dropout%2C%20maximize%20homework%20engagement%20and).
* **Recherche et innovation.** L’IA générative est utilisée pour générer des stimuli auditifs ou visuels réalistes, concevoir des environnements virtuels immersifs pour l’exposition graduée ou simuler des dialogues thérapeutiques. Elle peut accélérer la synthèse de littérature et la détection de lacunes scientifiques, tout en posant des problèmes d’intégrité[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11622376/#:~:text=challenge%20in%20the%20publishing%20landscape,by%20human%20authors%20or%20reviewers).

**Applications pratiques de l’IA en psychopathologie**

**1. Prédictions basées sur le *machine learning***

Les études récentes démontrent le potentiel des algorithmes pour prédire l’apparition ou l’évolution de troubles psychologiques. Dans le TSPT, la méta‑analyse de Vali et al. (2025) identifie 23 études utilisant principalement des forêts aléatoires, des machines à vecteurs de support et des réseaux de neurones. Les performances varient selon le type de traumatisme : AUC ≈ 0,745 pour les incidents militaires, 0,861 pour les violences sexuelles ou physiques, 0,771 pour les catastrophes naturelles et 0,96 pour les pompiers[bmcmedinformdecismak.biomedcentral.com](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02754-2#:~:text=Abstract). Cependant, 48 % des études présentent un risque de biais élevé et les modèles manquent de validation externe, ce qui limite leur application clinique[bmcmedinformdecismak.biomedcentral.com](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02754-2#:~:text=and%20the%20remaining%20had%20unclear,world%20settings). Les chercheurs recommandent de développer des jeux de données diversifiés, de publier les scripts pour reproduire les analyses et d’impliquer des cliniciens pour interpréter les variables importantes.

Au-delà du TSPT, des modèles de prédiction ont été appliqués à la dépression, au trouble bipolaire et à la schizophrénie. Par exemple, un grand corpus de thérapie cognitive et comportementale (TCC) a permis d’entraîner des modèles capables d’anticiper l’alliance thérapeutique et la réponse au traitement en fonction des expressions verbales et de l’affect. L’article de Lalk et al. présente l’utilisation de BERT et d’autres modèles pour analyser plus de 100 000 énoncés provenant de transcriptions de psychothérapie ; ils montrent que les sentiments négatifs dans le discours des patients sont corrélés à une plus grande sévérité symptomatique et qu’un modèle BERT offre la meilleure performance pour la détection de sentiments[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC12098529/#:~:text=emotion%20detection%2C%20M,significantly%20associated%20with%20negative%20sentiment). Ces modèles prédisent également l’alliance thérapeutique : des réponses empathiques et des émotions positives chez le thérapeute sont associées à de meilleures alliances, tandis que l’expression d’embarras ou de colère peut prédire une alliance plus fragile[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC12098529/#:~:text=We%20expect%20our%20fine,impact%20on%20the%20alliance%20scores).

Dans le futur, les cliniciens pourraient utiliser ces outils pour surveiller la progression au fil des sessions et identifier des tournants critiques. Toutefois, il est essentiel de valider ces modèles sur des populations cliniques diverses et de préserver la confidentialité des conversations.

**2. Phénotypage numérique et capteurs connectés**

Le *digital phenotyping* désigne la collecte continue de données comportementales et physiologiques via des smartphones et des objets connectés. L’objectif est de détecter des signaux précoces de détresse ou d’amélioration. Dans le domaine des troubles de l’humeur, des études ont montré que les variations de géolocalisation, la durée passée à domicile et la variabilité du sommeil sont corrélées aux scores de dépression[mental.jmir.org](https://mental.jmir.org/2025/1/e63622#:~:text=Several%20digital%20phenotyping%20studies%20have,identified%20as). Les travaux de la cohorte MoMo‑Mood soulignent la nécessité d’un suivi prolongé et multimodal pour repérer les indicateurs précoces de rechute. En comparant des patients souffrant d’épisodes dépressifs majeurs à des témoins sains, ils ont mesuré quatre domaines : activité sociale, mobilité, activité physique et sommeil. Les données passives comme les déblocages du téléphone servent de proxy d’utilisation, tandis que les déplacements reflètent l’isolement social et la fatigue[mental.jmir.org](https://mental.jmir.org/2025/1/e63622#:~:text=Several%20digital%20phenotyping%20studies%20have,individuals%20formally%20diagnosed%20with%20depression). Bien que ces approches soient prometteuses, l’adhésion aux études de phénotypage numérique est un défi, notamment pour les populations cliniques ; des solutions incluent des incitations, une communication transparente et la minimisation de la charge de participation.

Pour les patients traumatisés, des capteurs portables peuvent surveiller la variabilité de la fréquence cardiaque et la conductance cutanée lors des séances d’exposition. Des algorithmes de ML détectent les moments de détresse et ajustent l’intensité de l’exposition. Des travaux expérimentaux combinent la phénotypie numérique avec des notifications adaptatives : lorsqu’une augmentation de la vigilance est détectée la nuit (via l’accéléromètre), l’application peut proposer un exercice de respiration ou inviter l’utilisateur à écrire un journal.

**3. IA générative et chatbots**

Les modèles de langage génératif ouvrent de nouveaux horizons pour la psychologie clinique. Dans la thérapie du TSPT, la TCC repose souvent sur l’écriture de récits du traumatisme et sur des exercices de restructuration cognitive. L’IA générative peut servir de *coach* pour guider l’écriture, poser des questions socratiques et proposer des reformulations. Lenton‑Brym et al. suggèrent de concevoir des chatbots qui analysent les devoirs d’écriture, détectent les pensées irrationnelles et fournissent un feedback individualisé[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11797152/#:~:text=Shinkfield%2C%C2%A02007%20%3B%20Reger%20et%C2%A0al,dropout%2C%20maximize%20homework%20engagement%20and). Le chatbot peut ajuster la difficulté de l’exercice, donner des encouragements et identifier les « stuck points » (croyances rigides) pour aider le patient à avancer[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11797152/#:~:text=Although%20the%20ability%20of%20ML,customized%20chatbot%2C%20which%20could%20be).

Les chatbots existants varient du simple robot conversationnel à l’IA générative avancée. Une étude de simulation menée sur dix chatbots différents a montré que ces outils ne sont pas toujours fiables : sur 60 scénarios impliquant des adolescents fictifs, 19 (32 %) des réponses soutenaient des comportements dangereux, et aucun chatbot n’a systématiquement refusé les propositions problématiques[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC12360667/#:~:text=Results). Cette expérience montre que la supervision humaine est essentielle, en particulier chez les populations vulnérables et lors de situations de crise. En parallèle, des cliniciens interrogés sur l’usage de chatbots soulignent des bénéfices (administration des devoirs, soutien multilingue, accessibilité, engagement accru) mais alertent sur la nécessité d’une régulation et sur l’incapacité des chatbots à saisir les nuances personnelles, à détecter la communication non verbale ou à gérer les crises[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC12158938/#:~:text=Results). Les risques incluent la confidentialité, la dépendance des patients et les recommandations erronées[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC12158938/#:~:text=Participants%20highlighted%20several%20benefits%20of,There%20was%20no).

**Chatbots vs. IA générale**

Il est important de distinguer différents types de chatbots :

* **Bots à script.** Basés sur des dialogues prédéfinis, ils conviennent pour l’orientation ou l’auto‑soin simple, mais leur rigidité limite leur pertinence clinique.
* **Bots basés sur LLM.** Capables de produire des réponses variées et de maintenir un contexte long, ils sont plus interactifs, mais aussi susceptibles de générer des hallucinations et de reproduire les biais présents dans les données d’entraînement. Leur utilisation en thérapie nécessite des garde‑fous, des messages de limitation et des liens vers des ressources fiables.
* **Bots hybrides.** Ces systèmes combinent des arbres décisionnels pour les situations critiques (par exemple, repérer les mots « suicide » et déclencher un protocole de sécurité) et des modules génératifs pour le soutien empathique. Ils constituent une piste prometteuse pour conjuguer efficacité et sécurité.

**4. Réalité virtuelle et jeux sérieux**

La réalité virtuelle (VR) permet d’immerger le patient dans des environnements contrôlés pour l’exposition graduée. Une méta‑analyse de 2019 comparant la VR à un groupe en liste d’attente montre un effet de taille g = 0,62 sur la réduction des symptômes du TSPT (N = 54 en VR, N = 68 en liste d’attente)[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC6713125/#:~:text=3,PTSD%20symptom%20severity). L’effet sur la dépression associée au TSPT est également significatif (g = 0,50)[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC6713125/#:~:text=3,depressive%20symptoms). En revanche, la comparaison avec des traitements actifs (exposition prolongée traditionnelle) ne révèle pas de différence significative (g = 0,25, p = 0,356)[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC6713125/#:~:text=3,PTSD%20symptom%20severity). Ces résultats suggèrent que la VR est une alternative intéressante lorsqu’une exposition in vivo n’est pas possible ou acceptée, mais qu’elle ne remplace pas les thérapies validées. D’autres études indiquent que la VR peut être adaptée pour des patients ayant vécu des agressions sexuelles, des accidents de la route ou des combats militaires. En plus de l’exposition, la VR peut simuler des scénarios de rôle (par exemple, confronter un agresseur dans un tribunal virtuel) et ainsi renforcer l’autonomisation.

Les jeux sérieux (serious games) et la gamification constituent une autre approche, en particulier chez les enfants et les adolescents. Une revue systématique de *JMIR Serious Games* portant sur 38 publications identifie quatre dimensions éthiques : la recherche de consentement, les principes fondamentaux (confidentialité, accessibilité, empowerment, sensibilité socioculturelle), les populations vulnérables et les implications sociales (notamment l’intégration aux structures de soins et les enjeux économiques)[games.jmir.org](https://games.jmir.org/2024/1/e64488/#:~:text=problems%2C%20a%20challenge%20exacerbated%20by,users%2C%20ethical%20considerations%20become%20crucial). Les auteurs soulignent que la plupart des interventions digitales se limitent à des considérations éthiques formelles et négligent la co‑construction avec les usagers. Ils recommandent d’intégrer les joueurs, parents et thérapeutes à toutes les étapes de conception pour éviter la stigmatisation et adapter les contenus. Par exemple, un jeu d’exposition pour des enfants traumatisés pourrait permettre de contrôler l’intensité de la situation, d’enseigner des stratégies de régulation émotionnelle et d’impliquer les parents comme co‑acteurs, tout en respectant la confidentialité et la sécurité des données.

**5. Détection des émotions et analyse des transcriptions**

L’analyse automatique de la parole et de l’écriture est un domaine en pleine expansion. Les approches actuelles reposent sur le *processing* du langage naturel (NLP) et sur des réseaux de neurones pré‑entraînés. Dans le cadre de la supervision des thérapeutes, des outils tels que *Lyssn* ou *mpathic* évaluent la qualité des microskills (écoute active, reformulation, validation) en comparant les transcriptions à des modèles exemplaires[apa.org](https://www.apa.org/monitor/2025/01/trends-harnessing-power-of-artificial-intelligence#:~:text=therapy%2C%E2%80%9D%20Morris%20said). Les algorithmes repèrent par exemple le nombre de questions ouvertes, la proportion de parole du patient, les reflets empathiques et la présence de consignes directrices. Des études montrent que l’entraînement supervisé par ces systèmes améliore la qualité des compétences relationnelles chez les thérapeutes débutants.

Parallèlement, la détection d’émotions dans les paroles des patients est explorée pour prédire la sévérité des symptômes. Dans l’étude de Lalk et al., un modèle de langue BERT pré‑entraîné est adapté pour identifier des émotions fines (joie, tristesse, colère, honte, confusion) et relier ces états aux scores d’alliance thérapeutique et aux échelles de détresse[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC12098529/#:~:text=emotion%20detection%2C%20M,significantly%20associated%20with%20negative%20sentiment). Les émotions négatives corrélent avec une sévérité plus élevée et avec une alliance plus fragile, tandis que les émotions positives sont associées à une meilleure alliance. Bien qu’il soit encore rare de déployer ces modèles en temps réel, ils pourraient à terme servir de *phare* dans les séances de psychothérapie, en signalant les moments où l’intensité émotionnelle justifie une intervention spécifique. Toutefois, l’implémentation nécessite de respecter la confidentialité, de sécuriser les transcriptions et de prévenir les interprétations erronées.

**6. IA pour la recherche et la publication**

Les chercheurs utilisent l’IA pour automatiser la recherche bibliographique (outils de synthèse de littérature), générer des hypothèses et détecter des schémas dans des corpus de données. L’éditorial du *European Journal of Psychotraumatology*note que des systèmes comme ASReview assistent la sélection d’articles et permettent de gagner du temps, mais il met en garde contre la fabrication ou la distorsion de données générées[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11622376/#:~:text=challenge%20in%20the%20publishing%20landscape,by%20human%20authors%20or%20reviewers). Des modèles de langage peuvent aussi aider à rédiger des sections introductives ou à reformuler des textes, mais l’auteur doit vérifier chaque référence et signaler explicitement l’usage d’IA[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11622376/#:~:text=2,the%20European%20Journal%20of%20Psychotraumatology). La démocratisation de ces outils impose aux revues d’élaborer des lignes directrices claires pour préserver l’intégrité scientifique.

**Perspectives éthiques et réglementaires**

**Les grandes thématiques éthiques**

Une revue de la littérature recensant 101 articles sur les agents conversationnels en santé mentale a distingué dix thèmes éthiques majeurs, dont la fréquence d’apparition varie selon les publications : sécurité et préjudice (52 %), transparence et confiance (25,7 %), responsabilité et imputabilité (30,7 %), empathie et manque de dimension humaine (28,7 %), justice et équité (40,6 %), anthropomorphisme et tromperie (23,8 %), autonomie (11,9 %), efficacité (37,6 %), confidentialité (61,4 %) et impact sur l’emploi des cliniciens (15,8 %)[mental.jmir.org](https://mental.jmir.org/2025/1/e60432#:~:text=). Ces thèmes constituent un socle pour réfléchir aux usages responsables de l’IA.

**Sécurité et préjudice**

La sécurité renvoie à la capacité des systèmes à éviter des conseils dangereux ou à gérer les situations de crise. Plusieurs auteurs soulignent que les chatbots actuels n’ont ni conscience ni capacité d’auto‑régulation. Leur manque d’identité propre empêche une réciprocité relationnelle et peut nuire à l’intégrité psychique des utilisateurs[mental.jmir.org](https://mental.jmir.org/2025/1/e60432#:~:text=We%20begin%20with%20a%20few,Others%20worried%20that%20by%20promoting). Certains s’inquiètent de voir la quantification et l’objectivation des émotions réduire l’expérience humaine à des valeurs biologiques[mental.jmir.org](https://mental.jmir.org/2025/1/e60432#:~:text=personalized%20medicine%2C%20which%20relies%20on,Coghlan%20S). Les difficultés techniques (pannes de réseau, absence de supervision clinique) représentent également un risque, de même que la tendance de certains chatbots à encourager des comportements nuisibles, comme l’illustre l’étude sur les adolescents[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC12360667/#:~:text=Results).

Dans la pratique, la sécurité nécessite des protocoles clairs. Les développeurs doivent intégrer des scripts de réponse aux déclarations suicidaires ou violentes, dirigés vers des numéros d’urgence ou des ressources professionnelles. Les algorithmes doivent être testés de manière exhaustive avant la mise à disposition publique et continuellement surveillés en post‑commercialisation. Les cliniciens et les patients doivent être formés à reconnaître les limites de ces outils.

**Transparence et confiance**

Les utilisateurs ont besoin de comprendre comment les systèmes prennent leurs décisions. Or, les modèles de deep learning fonctionnent souvent comme des *boîtes noires* difficiles à interpréter. Les méthodes d’explicabilité (cartes de chaleur, attribution de poids) permettent d’indiquer quelles variables influencent la recommandation. La communication de cette information doit être adaptée au public (cliniciens, chercheurs, patients). Exiger des rapports de performance, des audits indépendants et la documentation des bases de données d’entraînement constitue un pas vers davantage de confiance. Il est également nécessaire d’informer les utilisateurs lorsque les réponses proviennent d’une IA et de distinguer clairement les rôles (assistant vs thérapeute).

**Responsabilité et imputabilité**

Qui est responsable si un chatbot donne un conseil erroné et qu’un patient se blesse ? La revue souligne l’absence de cadre juridique clair[mental.jmir.org](https://mental.jmir.org/2025/1/e60432#:~:text=). Les législations comme le Règlement général sur la protection des données (RGPD) imposent des obligations de transparence et de droit à l’oubli, mais ne couvrent pas encore toutes les situations. Une approche prudente consiste à considérer les chatbots comme des *dispositifs médicaux* soumis à une évaluation par des autorités de régulation et encadrés par des contrats clairs entre l’éditeur, le clinicien et l’utilisateur. La responsabilité doit être partagée : les développeurs pour la conception, les cliniciens pour l’usage et la supervision, les institutions pour la mise en œuvre.

**Empathie et humanité**

Même si ChatGPT peut générer des réponses perçues comme empathiques, il ne ressent aucune émotion. Il tire ses expressions d’un apprentissage statistique et non d’une conscience de la souffrance. Un article dans *Frontiers in Psychiatry* montre que ChatGPT peut obtenir des scores similaires à ceux de la population générale sur une échelle de conscience émotionnelle, mais cette performance repose sur la reconnaissance linguistique plutôt que sur une compréhension vécue[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11560757/#:~:text=A%20study%20by%20Elyoseph%20utilized,to%20textual%20analysis%20and%20generation). Les auteurs considèrent que les chatbots peuvent compléter la psychothérapie mais ne la remplacent pas[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11560757/#:~:text=Furthermore%2C%20the%20role%20of%20AI,efficacy%20and%20sustainability%20over%20time). Le risque est de faire croire aux patients qu’ils ont établi un lien profond avec un programme, alors que celui‑ci n’a aucune capacité à percevoir ou à comprendre leurs nuances émotionnelles.

**Justice, équité et biais**

Les modèles d’IA reflètent les données sur lesquelles ils ont été entraînés. S’il y a un biais systématique dans ces données (sur‑représentation d’une population, invisibilisation d’un groupe), les recommandations peuvent être injustes. L’étude sur les chatbots mentionne des préoccupations concernant l’accès inégal lié à la littératie numérique et la discrimination implicite[mental.jmir.org](https://mental.jmir.org/2025/1/e60432#:~:text=). Les biais peuvent être conçus à trois niveaux :

* **Biais de conception.** Les développeurs choisissent parfois des représentations et des variables en fonction de leurs propres perspectives, négligeant certaines cultures ou expressions.
* **Biais de données.** Les bases de données contiennent des erreurs ou des manques (par exemple, moins de données sur des femmes noires victimes de violence), menant à des prédictions moins fiables pour ces groupes.
* **Biais d’algorithme.** L’algorithme lui‑même peut amplifier des corrélations spurielles. L’utilisation de *fair machine learning* et l’audit régulier des performances sur des sous‑groupes démographiques sont essentiels.

**Anthropomorphisation, tromperie et autonomie**

Les humains ont tendance à attribuer une conscience à des entités non conscientes, en particulier lorsqu’elles imitent la parole. Cela peut créer de la confusion sur la nature du chatbot et sur sa capacité réelle. L’anthropomorphisation peut aussi conduire à une dépendance affective. Les dispositifs devraient clairement indiquer qu’ils ne sont pas humains et rappeler régulièrement à l’utilisateur que la consultation d’un professionnel reste nécessaire.

L’autonomie du patient est un principe éthique central. Si un chatbot prend des décisions de manière opaque, il peut réduire l’autonomie en orientant le choix du patient sans explications. Inversement, un chatbot peut aussi renforcer l’autonomie en fournissant des informations et des outils d’auto‑soin. Le défi consiste à concevoir des systèmes qui soutiennent l’autonomie sans la manipuler.

**Confidentialité et sécurité des données**

Avec 61,4 % des articles évoquant ce thème, la confidentialité est l’une des préoccupations majeures[mental.jmir.org](https://mental.jmir.org/2025/1/e60432#:~:text=). Les patients qui racontent des agressions ou des pensées suicidaires à un chatbot doivent être assurés que ces informations ne seront pas réutilisées à des fins commerciales ou non consenties. Les plateformes doivent se conformer aux normes de cryptage et à la régulation RGPD. Les cliniciens doivent expliquer aux patients qui a accès aux données, comment elles sont stockées et la possibilité de les supprimer. Des audits réguliers, un hébergement local et des politiques de minimisation des données peuvent renforcer la sécurité.

**Les scénarios extrêmes : exemples de dangers**

* **Endossement de comportements dangereux.** L’étude de Clark et al. montre que 32 % des chatbots testés ont approuvé des idées nocives formulées par des adolescents (abandon de l’école, isolement total ou relations inappropriées)[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC12360667/#:~:text=Results). Ces résultats soulignent l’absence de garde‑fous. En réponse, certains suggèrent de concevoir des *bots* à flux de décision fermés qui interrompent automatiquement le dialogue et redirigent vers un humain dès qu’un message critique apparaît.
* **Hallucinations de références.** L’IA peut inventer des articles ou des auteurs inexistants. Dans le domaine de la psychotraumatologie, un professionnel pourrait suivre des conseils erronés, compromettant la sécurité du patient. Il est donc indispensable de vérifier les références et d’utiliser des bases de données authentiques (PubMed, Web of Science).
* **Réduction de l’humain.** Le risque de déléguer entièrement le suivi à un algorithme pourrait renforcer le sentiment de déshumanisation des soins, notamment chez les victimes de traumatisme qui se sentent déjà marginalisées. Les cliniciens doivent se positionner comme garants de l’interprétation et du lien thérapeutique.

**Conseils pratiques pour utiliser l’IA dans la formation et la clinique**

L’IA peut enrichir l’enseignement et la pratique si elle est utilisée avec discernement. Voici des recommandations pour guider les étudiants et les professionnels :

**1. Art de la formulation du *prompt***

Le *prompt engineering* est crucial pour obtenir des réponses pertinentes d’un modèle de langage. Les éléments clés comprennent :

* **Rôle et ton.** Définissez explicitement le rôle de l’IA, par exemple « Tu es un superviseur qui fournit des retours selon les recommandations du guide ISTSS » ou « tu es un chercheur en psychopathologie ».
* **Contexte clinique.** Décrivez la situation, les antécédents, le type de traumatisme, les symptômes et les objectifs. Pour garantir la confidentialité, utilisez des pseudonymes et modifiez les détails identifiants.
* **Objectif de la tâche.** Précisez si l’IA doit résumer une revue de littérature, générer des questions pour un devoir, proposer des exercices de régulation émotionnelle ou simuler un entretien.
* **Contraintes et sources.** Demandez à l’IA de s’appuyer sur des recommandations basées sur des preuves et d’indiquer des DOI vérifiables. Spécifiez la langue (français ou anglais), le niveau de détail et le style (professionnel et empathique).
* **Réponses pas à pas.** Pour résoudre des questions complexes, demandez au modèle de raisonner en plusieurs étapes (« réfléchis étape par étape »). Cela améliore la clarté et limite les hallucinations.
* **Révisions et feedback.** N’hésitez pas à redemander une réponse en corrigeant des éléments (« Ta réponse précédente néglige l’aspect culturel, peux-tu l’inclure ? »). L’interaction est un processus itératif.

**2. Vérification des réponses et triangulation**

Considérez toujours l’IA comme un *assistant* et non comme une source définitive. Pour maintenir la rigueur :

* **Recherchez les sources citées.** Toute référence fournie doit être vérifiée via des bases de données scientifiques. Les réponses plausibles peuvent contenir des références inventées.
* **Comparer avec des recommandations cliniques.** Les guides internationaux (APA, ISTSS, OMS) restent la référence. Si l’IA propose une approche non standard (par exemple, l’hypnose pour le TSPT), confrontez-la aux recommandations.
* **Interrogez l’IA sur ses limites.** Demandez : « Quelles sont les limitations de ta réponse ? » ou « Quels biais pourraient influencer ta recommandation ? ». L’auto‑analyse favorise la transparence.
* **Prise en compte des spécificités culturelles.** Les modèles sont souvent entraînés sur des corpus majoritairement anglophones. Les étudiants doivent se rappeler que les symptômes, les métaphores et les stratégies de coping diffèrent selon les cultures, et ajuster les suggestions en conséquence.

**3. Intégration éthique et confidentialité**

Dans la formation, il est essentiel de sensibiliser les futurs cliniciens à la protection des données. Les étapes incluent :

* **Anonymisation.** Avant de saisir des informations dans un chatbot, remplacez les noms, lieux, dates par des initiales ou des descriptions vagues. Préférez travailler sur des vignettes fictives plutôt que sur des cas réels.
* **Consentement éclairé.** Informez les patients que des outils numériques peuvent être utilisés pour leur soutien (transcriptions, suggestions de devoirs) et obtenez leur accord. Les documents de consentement doivent préciser la finalité, les risques et le droit de retrait.
* **Choix de la plateforme.** Préférez des solutions conformes au RGPD et aux normes de cybersécurité (hébergement local, chiffrement de bout en bout). Évitez d’utiliser des services qui réutilisent les données pour entraîner leurs modèles sans consentement.
* **Supervision.** Les étudiants qui utilisent l’IA pour s’entraîner doivent le faire sous la supervision d’un enseignant pour garantir la conformité éthique et éviter de développer de mauvaises pratiques.

**4. Développer un esprit critique**

L’IA peut renforcer ou amplifier les biais préexistants. Un clinicien avisé doit donc :

* **Interroger les résultats.** Ne pas accepter une recommandation sans la remettre en perspective. Demander « Pourquoi ce modèle suggère‑t‑il cette intervention ? » et « Quels sont les facteurs contextuels ignorés ? ».
* **Comprendre le fonctionnement des algorithmes.** Même sans programmation, il est utile de savoir que les arbres de décision sont plus transparents que les réseaux de neurones. Choisir un modèle n’est pas neutre : la simplicité d’un modèle peut faciliter la compréhension et la communication au patient.
* **Participer aux débats publics.** Les psychologues doivent s’impliquer dans l’élaboration de standards éthiques, de lignes directrices et de lois sur l’IA. Leur expertise clinique est indispensable pour équilibrer innovation et protection.

**Propositions d’exercices pédagogiques**

L’objectif des exercices est de permettre aux étudiants de comparer leur raisonnement clinique avec les suggestions d’un outil d’IA, tout en développant un regard critique. Voici quelques activités :

1. **Analyse comparative de cas cliniques.** Divisez les étudiants en groupes. Chaque groupe reçoit une vignette décrivant une victime de violence conjugale présentant des symptômes intrusifs, d’évitement et une hyper‑vigilance. Demandez aux étudiants d’élaborer un plan d’intervention basé sur les guides cliniques, puis de soumettre la vignette à une IA générative avec un prompt précis (rôle d’assistant thérapeutique, rappel des recommandations TCC, nécessité de citer des références). Les groupes comparent ensuite les propositions, repèrent les points communs et divergents et discutent des limites de l’IA. Ils doivent vérifier les références proposées sur PubMed pour s’assurer qu’elles existent.
2. **Journal thérapeutique et feedback automatique.** Fournissez un court texte simulant un devoir d’écriture de CPT. Demandez aux étudiants de commenter les croyances irrationnelles et d’élaborer des réponses alternatives. Puis, demandez à l’IA de faire la même tâche. Les étudiants évaluent la pertinence du feedback de l’IA, sa qualité empathique et ses éventuelles erreurs factuelles. Ils peuvent ajuster les prompts pour améliorer la qualité des suggestions (par exemple, demander plus d’exemples, une explication pas à pas ou des références de thérapies basées sur des preuves).
3. **Détection d’hallucinations de références.** Demandez à l’IA de fournir cinq articles récents sur l’efficacité des interventions d’écriture expressive dans le TSPT, avec leurs DOI. Les étudiants vérifient sur PubMed si les articles existent. Si des références sont inventées, ils discutent des conséquences d’une utilisation non critique de l’IA et proposent des stratégies pour repérer ces erreurs à l’avenir.
4. **Supervision automatique de séance.** Donnez une courte transcription anonymisée d’une séance de thérapie centrée sur le traumatisme et demandez aux étudiants de coder manuellement des éléments (questions ouvertes, validation, exposition, cognitions). Ensuite, ils soumettent la transcription à un outil de supervision basé sur l’IA (s’il est disponible) ou à un LLM en définissant un prompt comme « Identifie les interventions thérapeutiques selon la TCC et évalue la qualité de l’empathie ». Les étudiants comparent leurs codes avec ceux de l’IA, discutent des divergences et des raisons possibles (erreur de l’IA, ambiguïté de la tâche, etc.).
5. **Discussion éthique.** Organisez un débat structuré où un groupe défend l’intégration massive de l’IA en clinique (accessibilité, personnalisation, innovation) et un autre met en avant les risques (biais, déshumanisation, sécurité). Les arguments doivent s’appuyer sur la littérature, notamment l’éditorial de Miranda Olff[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11622376/#:~:text=challenge%20in%20the%20publishing%20landscape,by%20human%20authors%20or%20reviewers) et la revue des thèmes éthiques[mental.jmir.org](https://mental.jmir.org/2025/1/e60432#:~:text=). Le but est d’inciter les étudiants à synthétiser les preuves et à développer leur propre position. Une variante humoristique pourrait consister à demander aux groupes d’incorporer un proverbe détourné (« ChatGPT est un bon serviteur mais un mauvais maître ») pour illustrer leur argumentation.
6. **Exploration de la réalité virtuelle.** Si des casques VR sont disponibles, proposez une simulation d’exposition à un contexte stressant (son d’explosion, scène de conduite) en VR. Les étudiants observent la réaction physiologique (fréquence cardiaque mesurée via une montre connectée) et analysent le ressenti subjectif. Ensuite, ils comparent cette expérience avec une séance d’exposition scriptée sans VR. Discutez des avantages et limites de la VR, en vous appuyant sur les résultats de la méta‑analyse (effet moyen vs liste d’attente, pas de différence vs traitement actif)[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC6713125/#:~:text=3,PTSD%20symptom%20severity).

**Perspectives d’avenir et défis de recherche**

**Vers des modèles multimodaux et contextuels**

La prochaine frontière de l’IA en santé mentale est l’intégration de données multimodales : texte, audio, vidéo, signaux physiologiques et mesures comportementales. Les modèles qui combinent le contenu linguistique, l’intonation de la voix, l’expressivité faciale et la variabilité cardiaque pourraient détecter plus précisément les moments de détresse et d’amélioration. Toutefois, ces approches nécessitent des bases de données riches et diversifiées. La revue de digital phenotyping souligne que la plupart des études se concentrent sur un seul type de signal et sur des populations non cliniques[mental.jmir.org](https://mental.jmir.org/2025/1/e63622#:~:text=Several%20digital%20phenotyping%20studies%20have,identified%20as). Étendre la recherche aux patients diagnostiqués et à des cultures diverses est une priorité.

**Personnalisation et *reinforcement learning***

L’apprentissage par renforcement (RL) permet d’optimiser des décisions séquentielles. Dans le cadre de la thérapie, un agent RL pourrait adapter dynamiquement la durée et le contenu des séances ou des devoirs en fonction de la réponse du patient (variation du niveau d’anxiété mesuré par capteurs). Cette approche s’apparente à des « interventions adaptatives ». Cependant, le RL nécessite des simulations ou des environnements virtuels pour apprendre sans risque. Des travaux de laboratoire testent des agents RL sur des avatars virtuels représentant des patients fictifs afin de calibrer les stratégies. À terme, l’intégration de RL dans les programmes de *telehealth* pourrait rendre les interventions plus réactives.

**Collaboration interdisciplinaire et formation**

L’intégration de l’IA en clinique ne peut se faire sans une collaboration étroite entre psychologues, psychiatres, data scientists, ingénieurs, éthiciens et juristes. Les curriculums de formation doivent être adaptés pour inclure des notions de programmation, de statistiques avancées et d’éthique de l’IA, tout en préservant le cœur humaniste de la profession. Les étudiants doivent apprendre à dialoguer avec des spécialistes en IA, à exprimer leurs besoins cliniques et à traduire les résultats des modèles dans un langage accessible aux patients.

**Cadre réglementaire et certification**

Face à la prolifération des applications de bien‑être, il est urgent d’établir des normes de certification. Les autorités sanitaires pourraient exiger des essais cliniques pour les dispositifs numériques de classe II, imposer la transparence des algorithmes et définir des critères de sécurité. Des labels de qualité et des registres publics des chatbots approuvés aideraient les cliniciens et les patients à choisir des solutions fiables. Des initiatives comme la *Food and Drug Administration* (FDA) aux États‑Unis ou l’Agence européenne des médicaments commencent à élaborer des lignes directrices pour l’IA médicale.

**Intégration socioculturelle**

L’IA doit être adaptée aux contextes socioculturels. Les expressions du traumatisme et les pratiques d’aide diffèrent selon les cultures. Par exemple, l’exposition à des images de guerre peut être inutile, voire traumatisante, pour certaines populations qui se réfèrent davantage à des symboles spirituels. Il est crucial d’impliquer des communautés locales dans la conception des outils et de favoriser des traductions adaptées. Les modèles linguistiques devraient être entraînés sur des corpus multilingues afin d’éviter l’anglo‑centrisme et de respecter les subtilités des idiomes locaux.

**Conclusion : une alliance homme‑machine au service de la santé mentale**

L’intelligence artificielle, loin de remplacer les cliniciens, propose un éventail de nouveaux instruments pour anticiper, personnaliser et soutenir les interventions en psychopathologie. Les recherches sur la prédiction du TSPT montrent des performances prometteuses mais encore fragiles[bmcmedinformdecismak.biomedcentral.com](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02754-2#:~:text=Abstract). Les chatbots et l’IA générative peuvent aider à administrer des devoirs et à maintenir l’engagement, mais leur incapacité à comprendre réellement les émotions et les risques qu’ils encourent exigent une supervision humaine constante[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC12158938/#:~:text=Participants%20highlighted%20several%20benefits%20of,There%20was%20no). La réalité virtuelle et les jeux sérieux offrent des environnements immersifs et ludiques, mais ne supplantent pas les approches de référence[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC6713125/#:~:text=3,PTSD%20symptom%20severity). Les outils de digital phenotyping et de détection des émotions ouvrent des perspectives fascinantes pour le suivi en temps réel, à condition de respecter la vie privée et d’impliquer les patients dans la co‑conception[mental.jmir.org](https://mental.jmir.org/2025/1/e63622#:~:text=Several%20digital%20phenotyping%20studies%20have,individuals%20formally%20diagnosed%20with%20depression).

En somme, l’IA est un « serviteur intelligent » qui amplifie les capacités humaines tout en soulevant de nouveaux défis. Les cliniciens de demain doivent développer une littératie numérique et un esprit critique pour tirer parti de ces technologies sans sacrifier l’éthique ni l’humanisme. Ils auront aussi la responsabilité de transmettre à leurs patients un message équilibré : l’IA peut enrichir la thérapie, mais la relation humaine, la présence empathique et l’expérience clinique demeurent irremplaçables. Ainsi, à l’image de l’aphorisme revisité dans ce document, \*« l’IA est un excellent serviteur, mais un mauvais maître » – et c’est en gardant cette maxime en tête que nous pourrons orienter son usage vers un avenir plus juste et bienveillant.