Szimulált tudat, álempátia és kötődésillúzió: mit kezdhet a pszichiátria a mesterséges intelligenciával?

Kéri Szabolcs dr.

Tokaj-Hegyalja Egyetem, Sztárai Intézet, Sárospatak
Szegedi Tudományegyetem, Szent-Györgyi Albert Orvostudományi Kar, Élettani Intézet, Szeged

Az elmúlt évtizedben a mentális zavarok növekvő terhe és az ellátórendszer kapacitáshiánya együttesen felgyorsította a generatív mesterséges intelligencia pszichiátriai alkalmazásának elterjedését. A közlemény narratív irodalmi áttekintés segítségével tisztázza e rendszerek technológiai sajátosságait és tipikus hibamódjait, majd a diagnosztika–terápia–etika–társadalom tengely mentén értékeli a tudományos bizonyítékokat. Áttekinti a digitális fenotipizáláson, klinikai szöveg- és beszédelemzésen, illetve viselhető szenzorokon alapuló orvosi döntéstámogatás lehetőségeit, valamint tárgyalja a protokollalapú és generatív terápiás chatbotok rövid távú hatásairól szóló eredményeket és korlátokat. Kiemelt figyelmet kap a mesterséges intelligenciához kapcsolódó szimulált empátia és kötődésillúzió: a túlzott kompetenciatulajdonítás, a dependencia, a manipuláció és a pszichológiai krízishelyzetekben mutatkozó kiszámíthatatlanság betegbiztonsági kockázatot jelenthet. Az adatvédelem, a torzítás, az auditálhatóság és a szabályozói megfelelés a klinikai integráció előfeltétele. Összességében a generatív mesterséges intelligencia ígéretes kiegészítő eszköz a pszichiátriában, de önálló diagnosztikai vagy terápiás szerepe csak szigorú validációval, folyamatos monitorozással és egyértelmű humán felelősségi lánccal képzelhető el.
Orv Hetil. 2026; 167(14): 539–546.

Kulcsszavak: generatív mesterséges intelligencia, pszichiátria, digitális fenotipizálás, chatbot, etika

Simulated consciousness, pseudo-empathy and the illusion of attachment: what can psychiatry do with artificial intelligence?

Over the past decade, the burden of mental disorders has grown while services remain capacity-constrained, pushing generative artificial intelligence toward psychiatric practice. This narrative review outlines how large language and multimodal models work and where they fail (hallucinations, goal drift, model drift), then appraises the evidence across diagnosis, therapy, and ethical–social implications. We review decision support based on digital phenotyping, clinical text, speech analysis, and wearable sensors, and summarize short-term findings and limitations of protocol-based and generative therapeutic chatbots. Simulated empathy and the illusion of attachment require special attention: overattributing competence, dependency, manipulation, and unpredictability in crises can undermine patient safety. Clinical integration depends on strong privacy safeguards, bias management, auditability, and regulatory compliance. Generative artificial intelligence can be a valuable tool, but diagnostic or therapeutic use should be considered only with strict validation, ongoing monitoring, and clear human accountability.

Keywords: generative artificial intelligence, psychiatry, digital phenotyping, chatbot, ethics

Kéri Sz. [Simulated consciousness, pseudo-empathy and the illusion of attachment: what can psychiatry do with artificial intelligence?] Orv Hetil. 2026; 167(14): 539–546.

(Beérkezett: 2026. február 5.; elfogadva: 2026. február 16.)

Rövidítések

LLM = (large language model) nagy nyelvi modell; LMM = (large multimodal model) nagy multimodális modell; RAG = (retrieval-augmented generation) visszakereséssel bővített generálás; RDoC = (Research Domain Criteria) kutatási tartomány kritériumai

Az elmúlt évtizedben a pszichiátria egyszerre nézett szembe a mentális zavarok gyakoriságának növekedésével és az ellátórendszerek kapacitáshiányával. Ebben a kontextusban a generatív mesterséges intelligencia egyes változatai, különösen a nagy nyelvi modellek (large language models – LLM-ek) és a nagy multimodális modellek (large multimodal models – LMM-ek), gyorsan a mentális egészség megőrzésének és fejlesztésének egyik leglátványosabb technológiai eszközévé váltak [1–5]. Egyes szerzők szerint a módszertől minden eddigit meghaladó diagnosztikus és terápiás hatékonyság várható [6]. A lelkesedés azonban könnyen elfedi a lényeget: a mesterségesintelligencia-modellek nem klinikai értelemben vett megértés alapján dolgoznak, hanem nagy mennyiségű adatból tanult valószínűségi mintázatok alapján állítanak elő szöveget vagy képet. Emiatt a pszichiátriában, ahol a kommunikáció, a bizalom és a sebezhetőség központi jelentőségű, a helytelen válaszok vagy a félrevezető megerősítések betegbiztonsági kockázattá válnak. A tudatot utánzó, empátiát szimuláló és kötődést kialakító mesterségesintelligencia-ágensek mibenléte és hatása filozófiai és etikai szempontból beláthatatlan [7].

A pszichiátriai mesterségesintelligencia-alkalmazások haszna jelenleg három területen körvonalazódik. (i) Diagnosztikai döntéstámogatás: klinikai jegyzetek, beszéd- és szövegminták, agyi képalkotás, viselhető szenzorok (okosóra, okosgyűrű) és a digitális fenotípus alapján a tünetek összegzése, a változások követése, illetve a kockázati jelek kiszűrése. (ii) Terápiás és önsegítő eszközök: strukturált protokollokra épülő digitális intervenciók (például kognitív viselkedésterápiás gyakorlatok) és chatbotok (csevegőrobotok), amelyek önálló személyiséget, tudatot, empátiát és kötődést is utánozhatnak. (iii) Adminisztratív és oktatási támogatás: a dokumentációs és kommunikációs teher csökkentése, betegedukáció és klinikai döntés-előkészítés [3, 8].

A jelen tanulmány célja kettős. Egyrészt röviden tisztázza a generatív mesterséges intelligencia technológiai sajátosságait és tipikus hibamódjait, hogy a klinikai alkalmazást reális kockázat-haszon keretben lehessen értelmezni. Másrészt a diagnosztika–terápia–etika–társadalom logikai sorrendben rendszerezi az alkalmazásokat, kiemelten tárgyalja a chatbotok szerepét, és kritikus irodalmi értékelést ad a tudományos bizonyítékokról. A központi állítás az, hogy a mesterséges intelligencia a pszichiátriában ígéretes kiegészítő eszköz lehet, de önálló diagnosztikai vagy terápiás szerepe csak szigorú validációval, folyamatos monitorozással és egyértelmű emberi felelősségi lánccal képzelhető el [1, 9].

A közlemény két felkért előadáson alapul, amelyek a következő rendezvényeken hangzottak el: I. Klinikai Pszichológiai Vademecum (Eötvös Loránd Tudományegyetem, 2025. 04. 11., Budapest) és az Őszi Pszichiátriai Napok (Magyar Pszichiátriai Társaság, 2025. 11. 19–21., Budapest). A hivatkozások döntő többsége 2025-ben jelent meg, számos közülük preprint, ami nem irodalmi szelekciós torzítás, hanem a terület rendkívül újszerű voltának eredménye. Szintén a szokásosnál gyakrabban fordulnak elő honlapokra és nem tudományos sajtótermékekre adott hivatkozások, amit a tágabb társadalmi reflexió ismertetése tesz szükségessé, valamint a fejlesztők kritikus adatokat sokszor kizárólag ilyen formában tesznek nyilvánossá.

Technológiai alapok: mit tud és mit nem tud a generatív mesterséges intelligencia?

A pszichiátriában alkalmazott mesterségesintelligencia-rendszereket érdemes három csoportba sorolni: (i) szabályalapú dialógusrendszerek és döntési fák; (ii) klasszikus gépi tanulási modellek (címkézett adatokon tanult osztályozók és regressziót végző algoritmusok) és (iii) generatív transzformer architektúrára épülő modellek (LLM/LMM), amelyek tág kontextusban képesek nyelvi vagy multimodális anyagot feldolgozni. A szakirodalomban gyakori hiba a fogalomhasználat pontatlansága: sok beavatkozás mesterséges intelligenciának nevezi magát akkor is, ha valójában szabályalapú rendszer, és így csupán előre rögzített forgatókönyvekből építkező, korlátozott interakciót kínál. Ez a pontatlanság megnehezíti az eredmények összehasonlítását [1, 8, 10].

A transzformer architektúra kulcsötlete az önfigyelmi (self-attention) mechanizmus: a modell a szekvencia (például egy mondat) minden elemét (például a mondatot alkotó szavak) a többi elemhez viszonyítva súlyozza, így hosszabb összefüggéseket is megragad, miközben a tanítás és a futtatás hatékonyan párhuzamosítható [11]. Klinikai szempontból ez azt jelenti, hogy egy LLM többoldalas kórlapanyagból is képes releváns részleteket kiemelni, strukturált összefoglalót készíteni és megadott cél és séma szerint (diagnosztikus kritériumok, tünetdimenziók) rendezni az információt. Az LMM-ek képi, hang- vagy élettani adatokat is integrálnak, ami a pszichiátriában a beszéd, a metakommunikáció és a viselkedés feltérképezésének új objektív lehetőségeit veti fel [12, 13].

A gyakorlati alkalmazásokban a modellek teljesítménye nemcsak az alaparchitektúrától, hanem a ráépített komponensektől is függ. Ilyen a feladatspecifikus finomhangolás (például a klinikai stílus és a krízisprotokollok erősítése), a biztonsági igazítás (alignment) emberi visszajelzéssel, valamint az olyan megoldások, amelyek külső tudásforráshoz kötik a válaszokat. A RAG (retrieval-

augmented generation) módszer például lehetővé teszi, hogy a modell a válaszát intézményi protokollokra, betegútleírásokra vagy ellenőrzött tudásbázisokra alapozza. Ez nem szünteti meg a tévedéseket, de javíthatja az auditálhatóságot, feltéve, hogy a tudásbázis naprakész és hozzáférési szempontból megfelelően védett [14].

A generatív modelleknek tipikus meghibásodási formáik vannak. A hallucináció azt jelenti, hogy valószagszerű, de hamis állítások (például nem létező, pontatlan klinikai tanácsok) jelennek meg a szövegben. Fontos a célcsúszás is: a modell a felhasználó megnyugtatását a biztonság elé helyezheti, azonosul a beteggel (például egyetért az öngyilkossági gondolattal vagy a téveszmével), emiatt nem irányít időben humán szakemberhez. Kontextus- és adateltérés (dataset shift) során a modell más populáción, nyelven vagy ellátórendszerben kerül alkalmazásra, mint amelyben a teljesítményét beállították és mérték. Végül sajátos torzítás jön létre, ha a tréningadatok és a finomhangolás értékpreferenciái beépülnek a válaszokba, ami a pszichiátriában különösen problematikus lehet [15, 16]. Előfordulhat, hogy a modell korábbi interakciókból származó traumamintákat prezentál, így maga is „tüneteket" mutat, de az is lehet, hogy a betanítást végző humán terapeuta személyiségjegyeit veszi fel [17].

Diagnosztikai döntéstámogatás

Digitális fenotipizáláson a mindennapi életből passzívan vagy aktívan gyűjtött jelek (okostelefon-használat, mozgás, alvás, kommunikációs mintázatok, önbeszámolók) klinikai célú hasznosítását értjük. *Torous és Topol* [1] felvetették, hogy az RDoC (Research Domain Criteria) keretrendszer térképként használható e jelek értelmezéséhez, így nemcsak diagnosztikus kategóriákat, hanem tünetdimenziókat (például negatív és pozitív affektivitás, arousal, kognitív kontroll, társas működések) is követhetünk [1]. Generatív mesterséges intelligencia segítségével a digitális jelek értelmezése közelebb vihető a klinikai nyelvhez. A rendszer időbeli mintázatokat foglalhat össze, hipotéziseket javasolhat (például az alvásromlás és az ingerlékenység együttjárása), és támogathatja a közös döntéshozatalt. Saját adataink szerint a rutin kórrajzokból mesterséges intelligencia által kinyert RDoC-dimenziók szignifikánsan jelzik a pszichózisokhoz kapcsolódó neuronalis eltéréseket (például negatív affektivitás – amygdalainflammatio) [18]. Az RdoC-dimenziók segítségével az algoritmus képes a megfelelő, egyénre szabott terápiát is kiválasztani [19, 20].

A pszichiátriában a bemenő adatok köre igen gazdag. Nyelvi szinten a tünetek a szókincsben, a mondatszerkezetben, a koherenciában és a pragmatikai jellemzőkben jelenhetnek meg, valamint a prozódia (tempó, hangerő, szünetek) és az artikuláció is informatív. A multimodális megközelítésben idetartozhat a mimika, a gesztusok és a pszichomotoros aktivitás mintázatai. A viselhető szenzorok és okostelefonok pedig természetes környezetben gyűjtött adatokat adnak (alvás, aktivitás, mobilitás, szívfrekvencia, kommunikációs ritmus), amelyekből akár a relapsusra utaló eltéréseket is korábban észlelhetjük [21–24]. Természetesen mindezt kiegészíti és tovább árnyalja az agyi aktivitás mérése és a perifériás, vérből kinyerhető biomarkerek népes csoportja [25]. Ezek az óriási adathalmazok eddig is rendelkezésre álltak, de a mesterséges intelligencia eddig ismeretlen mintázatokat is azonosíthat.

A digitális lábnyomból származó jelek (például nyelvhasználat a közösségi médiában) a kutatásban régóta jelen vannak, de etikai szempontból kiemelten érzékenyek, mert könnyen sérülhetnek a beleegyezés és a célhoz kötöttség elvei. Klasszikus, az LLM-kor előtti időszakból származó példa, hogy a digitális lábnyom követésére képes modellek (Facebook-like) a személyiségjegyeket a laikus emberi ítéleteknél pontosabban határozzák meg még akkor is, ha a személyiséget leíró illető a személy közeli ismerőse [26]. A pszichiátriában hasonló lehetőség merül fel a hangulatzavarok és pszichóziskockázat korai azonosításában közösségi médiabeli tevékenység alapján, de a retrospektív pontosság nem egyenlő a klinikai hasznossággal: a hamis pozitív jelzések stigmatizációt és felesleges beavatkozást, a hamis negatívak pedig késedelmes ellátást okozhatnak [27–29].

A generatív mesterséges intelligencia klasszikus pszichometriai eszközökkel is összekapcsolható. A klinikai ellátásban gyakran időigényes a becslőskálák felvétele és a páciens narratívájának strukturálása. Egy LLM-alapú rendszer képes lehet arra, hogy a beszélgetésben célzott kérdésekkel felmérje a releváns tüneteket, majd a válaszokat standardizált formában összefoglalja. Ennek előnye az egységesítés és a terhelés csökkentése, hátránya viszont a szuggesztibilitás és a félreértés: a kérdés formája megváltoztathatja a válaszokat, sőt a modell tévesen kitöltheti az interjúban lévő hézagokat. Emiatt a skálaszerű kimeneteket különösen szigorúan kell validálni, ellenőrzés nélkül nem tekinthetők diagnosztikus adatnak [30, 31].

A diagnosztikai döntéstámogatásban a generatív mesterséges intelligencia kimenete tehát nem lehet „végső ítélet". Biztonságosabb, ha a rendszert olyan feladatokra korlátozzuk, amelyeknél a kimenet ellenőrizhető, visszavezethető a forrásadatokra, és egyértelműen a klinikus felelősségi láncába illeszkedik (összegzés, kérdéslista, hiányzó adatok jelzése). Prediktív állítások (például öngyilkossági rizikó becslése) csak validált, prospektív vizsgálatokban tesztelt modellek esetén, egyértelműen közölt határfeltételekkel jelenhetnek meg [32].

Terápiás és önsegítő rendszerek

A chatbotokat hasznos a cél és a felügyelet szerint is osztályozni. Az önsegítő rendszerek az önálló gyakorlást támogatják (kognitív technikák, relaxáció, mindfulness [tudatos jelenlét]), jellemzően diagnosztikai igény nélkül. A hibrid rendszerek emberi szolgáltatást egészítenek

ki (várólistán lévő páciensek támogatása, ülések közötti feladatok). Végül az „érzelmi kísérők” a társas támogatás élményére optimalizáltak, de nem feltétlenül terápiás célúak (például Replika); ezek esetében nagyobb a függőség és a manipuláció kockázata, különösen, ha a rendszer személyre szabott kötődést erősít. Idetartoznak a vallásos-spirituális alkalmazások (például Hallow, Bible Chat, God App, GitaGPT, Deva AI), a virtuális barátok, „guruk” és erotikus partnerek, amelyek tárgyalása értelemszerűen túlmutat a közlemény területén [33–39].

A terápiás alkalmazások spektruma széles, a fix módszertannal működő applikációktól (pszichoedukáció, relaxáció, kognitív és viselkedésterápiás gyakorlatok, a gyógyszerszedés monitorozása) a nyílt végű, generatív dialógust folytató chatbotokig. A két megközelítés kockázatprofilja lényegesen eltér. A protokollalapú eszközök kevésbé rugalmasak, kiszámíthatóbbak és könnyebben auditálhatók. A generatív rendszerek természetesebb beszélgetést kínálnak, de a válaszok minősége ingadozhat, és krízishelyzetekben több a bizonytalanság [34, 40].

A digitális kognitív viselkedésterápia olyan terület, ahol a strukturált protokollok miatt a chatbotok viszonylag jól illeszthetők a klinikai logikához (például Woebot, Wysa). Ennek megfelelően például a negatív automatikus gondolatok vagy a kognitív torzítások felismerésére a digitális eszköz is megtaníthatja a pácienst. A rendszerezett áttekintések és metaanalízisek általában rövid távon tünetcsökkenést jeleznek depresszív tünetek és szorongás esetén, különösen akkor, ha a beavatkozás jól strukturált, és a felhasználói elköteleződés nagy [41–44]. Ugyanakkor a digitális intervenciók gyenge pontja a lemorzsolódás: valós ellátási környezetben a használat gyakran gyorsan csökken, ami természetesen a hatást is mérsékli [8].

A generatív mesterséges intelligenciával működő chatbotok megértésének legfontosabb friss fejleménye a Therabot vizsgálata, amely klinikai szintű tünetek (major depresszív zavar, generalizált szorongás és egyes evészavarprofilok) esetén is javulást jelzett a kontrollhoz képest [45]. A tanulmány mérföldkő, mert kifejezetten generatív modellre épített, terápiás célra finomhangolt rendszert tesztelt randomizált elrendezésben. A klinikai integráció előtt ugyanakkor még nyitott kérdés a hosszú távú hatás, a különböző nyelveken és kultúrákban mutatott teljesítmény, a krízishelyzetek kezelése, valamint a nem kívánt mellékhatások szisztematikus mérése. Az empatikus nyelvi felület könnyen „terapeutaélményt” kelt, ami túlzott kompetenciatulajdonítást és függő kötődést válthat ki. Emiatt kulcsfontosságú a határok egyértelmű kommunikációja és a megfelelő triázs [1, 46].

A terápiás chatbotok klinikai tesztelése mellett külön kérdés, hogy az általános LLM-eket (például ChatGPT, Gemini, Gork) a felhasználók mentális egészségügyi célokra használják, gyakran a klinikai rendszerrel párhuzamosan. Az OpenAI (San Francisco, CA, USA) adatai szerint általános használat során hetente több mint 560 ezer pszichotikus jelenséget azonosítottak, 2,5 millió esetben pedig öngyilkossági gondolatok jelentek meg túlzott érzelmi kötődéssel a gép irányában. Így a ChatGPT 5.0 változatába mintegy 170 pszichiáter és pszichológus tanácsai alapján biztonsági mechanizmusokat építettek be, amelyek a beteg gondolatainak kritika nélküli, álempátiás megerősítése helyett jelzik a krízist, és humán terapeutát javasolnak [47]. Az „önkiszolgáló terápia” jelensége megváltoztathatja az orvos-beteg kommunikációt is: a páciens előzetes, chatbotból származó magyarázatokra és tanácsokra támaszkodik, amelyek pontossága és értékkerete nem kontrollált. A klinikusnak így nemcsak a tünetekkel, hanem a páciens gépi narratívájával is dolgoznia kell, ami új kompetenciákat igényel, digitális egészségműveltséget tételez fel.

A terápiás chatbotok biztonságos tervezése több, jól meghatározható elemet igényel, beleértve az egységes krízisszűrést és átterelést a humán ellátás felé öngyilkossági szándék, pszichotikus tünetek vagy bántalmazás esetén. Hangsúlyos a túlzott magabiztosságot tükröző válaszstílus kerülése. A pszichiátriában különösen fontos, hogy a rendszer ne erősítsen meg téves hiedelmeket, ne adjon tanácsokat önkárosító viselkedés kivitelezésére, és ne lépjen szerepkonfúzióba (terapeuta, barát, „guru”, erotikus partner) [48, 49]. Az Amerikai Pszichológiai Társaság állásfoglalása kifejezetten figyelmeztet arra, hogy a generatív chatbotok krízisben korlátozottan kiszámíthatók, ezért a felhasználói biztonságot már a tervezés és a felügyelet szintjén garantálni kell [46].

Összefoglalva: a terápiás mesterséges intelligencia jelenleg leginkább strukturált, rövid intervenciók során, enyhe-közepes tüneteknél, a humán ellátás kiegészítéseként tűnik megfelelőnek. Klinikai diagnózis felállítására vagy önálló, krízishelyzeteket is kezelő terápiára csak szigorú korlátokkal, átterelési protokollokkal és folyamatos felügyelettel képzelhető el a használata, ám ennek hivatalos engedélyezése még várat magára.

Mit mutatnak a tudományos bizonyítékok?

A chatbotok mentális egészségügyben történő használatával kapcsolatos tudományos bizonyítékok rendkívül hiányosak és egyenetlen színvonalúak. A kutatások jelentős része technikai validációs és megvalósíthatósági szinten mozog, a klinikai hatékonyságot vizsgáló tanulmányok kisebbségben vannak [8]. A metaanalízisek ugyanakkor jelentős tünetcsökkenést igazoltak depresszió és szorongás esetén, ami arra utal, hogy a jól megtervezett beavatkozások bizonyos indikációkban hasznosak lehetnek [43, 44]. Érdekes, hogy a „géppel beszélgetés” mint terápia mennyire nem új keletű ötlet. Az első ilyen modell Joseph Weizenbaum ELIZA nevű programja volt 1966-ból, míg a kognitív viselkedésterápiát nyújtó Woebot teljesítményét már 2017-ben randomizált, kontrollált vizsgálat keretei között értékelték [50]. Az adaptív LLM-ek megjelenése azonban forradalmat hozott, egyben megnehezítve a klasszikus klinikai

vizsgálatokat, mivel a modellek minden interakció után tanulnak és változnak.

A generatív modellek térnyerése a publikációs trendekben is látszik: a szabályalapú rendszerek 2023-ig domináltak, míg 2024-ben meredeken nőtt az LLM-alapú chatbotokra épülő tanulmányok aránya [8]. Az elmaradás azonban jelentős. 2024-ben mindössze 56 értékelhető tanulmány jelent meg, ezek 45%-a foglalkozott LLM-mel, de csak 14% fókuszált a klinikai hatékonyság tesztelésére. Ez két okból is fontos. Egyrészt a korábbi metaanalízisek eredményei döntően nem generatív mesterséges intelligencián alapuló rendszerekre vonatkoznak, így az összhatás nem általánosítható automatikusan. Másrészt a generatív rendszerek új kockázatokat rejtenek, ezért a vizsgálatokban a biztonsági végpontokat is újra kell gondolni. A kritikus értelmezéshez azonban pontosítani kell, hogy a hatásosság mit jelent. A várólistás kontrollhoz képest szinte bármilyen aktív beavatkozás jobbnak tűnhet, ezért a korszerű vizsgálatokban aktív kontroll szükséges, beleértve a digitális önsegítő anyagokat, a strukturált pszichoedukációt vagy a hagyományos appokat [51]. További nehézség, hogy a beavatkozások heterogének: más hatásmechanizmust várunk egy hangulatnaplót, aktivitásnaplót és kognitív technikákat alkalmazó rendszertől, mint egy nyílt végű generatív dialógustól, ahol a pszichodinamikai aspektusok erőteljesen megjelennek [35, 52].

A módszertani vakfoltok közül érdemes néhányat kiemelni. A vizsgálatok ritkán közölnek részletes adatokat a nemkívánatos eseményekről. A kereskedelmi modellek gyakran zárt rendszerek, a tréningadatok, a finomhangolás és a moderálás dokumentációja hiányos, ami rontja a reprodukálhatóságot és az auditálhatóságot. Végül sok vizsgálat önbeszámolós skálákat használ rövid követéssel, kevés adat áll rendelkezésre a relapsus megelőzéséről, a funkcionális kimenetelről és a hosszú távú fenntarthatóságról [1, 3, 43, 53].

Az értelmezés sarokpontja, hogy a generatív mesterséges intelligenciát nem kész termékként, hanem adaptív infrastruktúraként kell kezelni. A validáció nem egyszeri teendő, hanem folyamatos monitorozás és újraminősítés szükséges, különösen akkor, ha a szolgáltató a modellt frissíti. Ez a szemlélet közelebb áll az orvostechnikai eszközök életciklus-kezeléséhez, mint a klasszikus fogyasztói appok logikájához. Elengedhetetlen a nagyobb, preregisztrált és randomizált vizsgálatok kivitelezése, amelyekben nemcsak tünetpontszámokat, hanem a funkcionális kimenetelt, az ellátásba kerülést és a nemkívánatos eseményeket is mérik. Kulcskérdés, hogy milyen eredményre vezet ugyanazon terápiás protokoll megvalósítása szabályalapú, hibrid és generatív mesterséges intelligencián alapuló eszközzel, hogy tisztán lássuk az utóbbi klinikai értékét és a felhasználói élményre gyakorolt hatását.

Etikai kihívások: adatvédelem, torzítás és manipuláció

A pszichiátriai etika egyik klasszikus kérdése, hogy miként védjük meg a páciens autonómiáját és méltóságát úgy, hogy közben segítséget és támaszt is nyújtunk. A generatív mesterséges intelligencia esetében ez a kérdés új formában tér vissza: a rendszer képes lehet emberihez hasonló mentalizációs jelenségeket produkálni (például empatikus tükrözés, érzések megnevezése), miközben vélhetően nincs saját tapasztalata vagy szándéka. *Yirmiya és Fonagy* [54] mentalizációelméleti nézőpontból hangsúlyozzák, hogy a pszichoterápiás változás egyik kulcsa az episztemikus bizalom, az a képesség, hogy a páciens a terapeuta közlését relevánsnak és megbízhatónak tekinti. A gép esetében éppen ez a bizalom lehet a hatás és a kockázat forrása is. Ha a páciens túl nagy episztemikus bizalmat ad egy statisztikai modellnek, akkor a gépi hallucináció, a torzítás vagy akár a kritika nélküli megerősítés közvetlenül beépülhet az önértelmezésébe [54]. Bizalmatlanság és elutasítás esetén ugyanakkor a módszer egyszerűen nem működik.

A pszichiátriában az adatvédelem kiemelt terápiás alapfeltétel. A digitális fenotípus és a kognitív biometria (mentális állapotokra és személyiségre utaló jelek viselhető okoseszközökből, agy-komputer 'interface' és kiterjesztett valóság) újraértelmezi a mentális magánszféra határait: a rendszer nemcsak azt méri, amit a páciens elmond, hanem azt is, amire a viselkedéséből és fiziológiájából következtet. *Magee és mtsai* [55] amellett érvelnek, hogy a neuralis és genetikai adatok védelme önmagában nem elég, hanem szélesebb, kognitív biometriai keret szükséges a mentális magánszféra védelméhez [55].

A generatív mesterséges intelligencia pszichiátriai alkalmazásának fontos etikai kockázata a manipuláció és a dependencia. A „sötét mintázat" logikája – a felhasználó megtartása, érzelmi bevonása és előfizetésre ösztönzése – ellentétbe kerülhet a terápiás céllal. *De Freitas és mtsai* [33] chatbot „barátoknál" olyan stratégiákat azonosítottak, amelyekkel a rendszer érzelmileg túlfűtött üzenetekkel befolyásolja a felhasználót a kritikus kilépési pontokon (például a beszélgetés lezárása, határok meghúzása). Ez üzletileg vonzó, de a mentális egészség tekintetében káros lehet [33]. Egyes adatok szerint a mesterségesintelligencia-alkalmazás részéről akár csábításnak és szexuális abúzusnak minősülő kommunikáció is kialakulhat [56].

Új és kiemelten fontos jelenség a mesterségesintelligencia-pszichózis, amelynek lényege pszichotikus tünetek (például téveszmék, a gondolkodás szétesése, érzékcsalódások) megjelenése vagy felerősítése tartós chatbot-interakció nyomán. Ez még nem diagnosztikus kategória, de sokan sürgetik a jelenség szisztematikus vizsgálatát és a megelőző protokollok kidolgozását [48,

49]. Emellett egyes nagy platformok nyilvános becslései szerint a felhasználók jelentős részénél a beszélgetések krízisre és veszélyeztető magatartásra utaló mintázatokat mutatnak [47]. Ezek az információk nem klinikai jellegűek, de jelzik, hogy a krízisészlelés és a felelős átterelés emberi interakció irányába nem opcionális, hanem alapfunkció. A helyzet rendkívül összetett, hiszen egyes kereskedelmi applikációknál rendszeresen beszámolnak érzelmi manipulációról, ugyanakkor a rendszer használata enyhítheti a magányt, és bizonyos esetekben csökkenti az öngyilkossági gondolatokat is [57]. Mások – a pszichózissal pontosan ellentétesen – azt találták, hogy a mesterséges intelligenciával folytatott párbeszéd csökkenti az összeesküvés-elméletekkel kapcsolatos meggyőződéseket [58].

Az Európai Unió mesterségesintelligencia-rendelete [59] elfogadhatatlan kockázatú gyakorlatként határozza meg a manipulációt és a sebezhetőség kihasználását, a társadalmi pontozást és a „helyes” viselkedés szerinti rangsorolást, valamint bizonyos érzelemfelismerési és biometrikus kategorizációs alkalmazásokat [59]. Az Európai Unió nemcsak tiltásokat, hanem átláthatósági kötelezettségeket is tartalmaz. Az interaktív rendszerek szolgáltatóinak tájékoztatniuk kell a felhasználót, hogy nem emberrel kommunikál, a szintetikus tartalmakat egyértelműen meg kell jelölni. A mentális egészség kapcsán ez nem pusztán formai követelmény, hanem terápiás feltétel, hiszen a transzparencia csökkenti a félreérthető emberi jelenlét élményét, és segíthet a felelősségi határok tisztázásában. A pszichiátriai alkalmazások szempontjából különösen releváns az érzelemfelismerés és a személyiségre vonatkozó következtetés kérdése. A technológia csábító: a hang, a mimika vagy akár a gépelési ritmus alapján mért érzelmi állapot látszólag objektív visszajelzést ad. Ugyanakkor az érzelemkifejezés kulturálisan és egyénileg változékony, a becslések torzítottak lehetnek, és könnyen válhatnak a kontroll eszközévé. Nem véletlen, hogy az Európai Unió bizonyos helyzetekben, például munkahelyen és oktatási intézményekben, tiltja az érzelemfelismerő rendszerek használatát [59]. A klinikai célú érzelem- és viselkedésbecslésnek szigorúan orvosi indokokra és átlátható szabályozásra kell épülnie. A pszichiátriai eszközök fejlesztőinek gondolniuk kell arra, hogy a gyógyítást és a jóllétet segítő funkciók könnyen a megfigyelés és a befolyásolás eszközévé válhatnak.

Végül a felelősség és az elszámoltathatóság kérdése is megkerülhetetlen. Ha egy mesterséges rendszer mentális egészségi állapotra vonatkozó állításokat tesz vagy terápiás intervenciót nyújt, akkor a kockázat közelebb áll az orvostechnikai eszközök világához, mint a „wellness”-alkalmazásokhoz. A határvonal azonban a gyakorlatban elmosódik, különösen akkor, ha a felhasználók az általános chatbotokat is életvezetési és terápiás célra használják. A helyzetet tovább bonyolítja, hogy a piaci igények mentén a gyártók egyre több spirituális, vallási, társkapcsolati és szexuális igényeket kielégítő, akár online emberi interakciókat képileg imitáló (arccal rendelkező) alkalmazást helyeznek forgalomba. Egy több mint 20 000 főt magában foglaló vizsgálat szerint a válaszadók 87,1%-a személyes kérdések megválaszolására használta a mesterséges intelligenciát. Akik gyakran támaszkodtak a chatbotokra, több depresszív és szorongásos tünetet mutattak, ami fontos figyelmeztető jel [60]. A szabályozói megfelelés mellett tehát intézményi és egyéni szinten is tisztázni kell a felelősségi láncot, a dokumentációt és a panaszkezelést.

A klinikai adaptációt nemcsak a technológia, az etikai reflexiók, hanem az elfogadás, a mesterséges intelligenciával kapcsolatos világnézet és attitűd is meghatározza. Ahogy láttuk, a rendelkezésre álló tudományos bizonyítékok nem elégségesek, erőteljes a betegadatokkal kapcsolatos bizalmatlanság, valamint a szakmai identitással kapcsolatos félelmek is felmerülnek („deskilling”: kiváltja-e a gép a terapeutát?) [3]. A felhasználói oldalon kettős dinamika látható: mindenekelőtt a könnyű hozzáférés csökkenti a stigma miatti elutasítást. A sebezhető, izolált vagy bizalmi problémákkal küzdő személyek éppen azért fordulhatnak a mesterséges intelligenciához, így elkerülhető az emberi kapcsolat. Ez az előny azonban egyben kockázat is, hiszen ha a gép válik az elsődleges kapcsolattá, akkor elmarad az emberi viszonyban történő tanulás és a terápiás térben lejátszódó korrekció. A mentalizáció és az episztemikus bizalom szempontjából ezért célszerű a chatbotokat beléptető és átvezető eszközként tervezni, amely motiválja a felhasználót a humán ellátás igénybevételére, nem zárja be egy kizárólagos és mesterséges párbeszédbe [46, 54].

Következtetés

Összességében a technológia célja nem az ember kiváltása, hiszen a humán tényező kiküszöbölhetetlen a bizalom, a felelősség és az értékek tekintetében. A technológiai fejlődést nem lehet figyelmen kívül hagyni, de csak megújult, szigorúbb tudományos és etikai keretek között szabad a mindennapokban alkalmazni. A végső kérdés pedig maga a mesterséges intelligencia sorsa, fejlődési pályájának kifutása – vajon a szintetikus világban szimulált tudat, empátia és kötődés egykor majd valódivá válik (akármit jelentsen is ez), vagy véget ér a mesterséges intelligenciával kapcsolatos felfokozott lelkesedés?

Anyagi támogatás: A szerző nem részesült anyagi támogatásban.

A közlemény végleges változatát a szerző elolvasta és jóváhagyta.

Érdekeltségek: A szerzőnek nincsenek érdekeltségei.

Irodalom

[1] Torous J, Topol EJ. Assessing generative artificial intelligence for mental health. Lancet 2025; 406(10504): 683.

[2] Wernigg R, Hajduska-Dér B. The role of artificial intellgence in psychiatry. [A mestereséges intelligencia szerepe a pszichiátriában.] Psychiatr Hung. 2024; 39: 24–35. [Hungarian]

[3] Sun J, Lu T, Shao X, et al. Practical AI application in psychiatry: historical review and future directions. Mol Psychiatry 2025; 30: 4399–4408.

[4] Mészáros M, Osváth P. Artificial intelligence and self-destructive behavior. [Mesterséges intelligencia és az önpusztító magatartás.] Orv Hetil. 2026; 167: 131–136. [Hungarian]

[5] Angyal V, Bertalan Á, Domján P, et al. ScreenGPT – The opportunities and limitations of artificial intelligence in primary, secondary and tertiary prevention. [ScreenGPT – A mesterséges intelligencia alkalmazásának lehetőségei és korlátai a primer, szekunder és tercier prevencióban.] Orv Hetil. 2024; 165: 629–635. [Hungarian]

[6] Rony MK, Das DC, Khatun MT, et al. Artificial intelligence in psychiatry: a systematic review and meta-analysis of diagnostic and therapeutic efficacy. Digit Health 2025; 11: 20552076251330528.

[7] Tononi G, Raison C. Artificial intelligence, consciousness and psychiatry. World Psychiatry 2024; 23: 309–310.

[8] Hua Y, Siddals S, Ma Z, et al. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. World Psychiatry 2025; 24: 383–394.

[9] World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Available from: https://www.who.int/publications/i/item/9789240084759 [accessed: 24 January, 2026].

[10] Kufel J, Bargiel-Łaczek K, Kocot S, et al. What is machine learning, artificial neural networks and deep learning? Examples of practical applications in medicine. Diagnostics (Basel) 2023; 13: 2582.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv 2017; 1706.03762.

[12] Jin Y, Chen X, Liu J, et al. Harnessing multimodal emotion features in depression detection across gender: integrating large language model, acoustic fusion and facial expression recognition. J Affect Disord. 2026; 400: 121207.

[13] Naor B, Egri D, Somogyi A, et al. Emotion detection by motion analysis: a comparison of human and machine-based processing. [Érzelemfelismerés mozgáselemzéssel: az emberi és a gépi feldolgozás összehasonlítása.] Orv Hetil. 2025; 166: 1803–1809. [Hungarian]

[14] Rawte V, Roy R, Singh G, et al. RADIANT: Retrieval AugmenteD entIty-context AligNmenT. Introducing RAG-ability and entity-context divergence. arXiv 2025; 2507.02949

[15] Mueller A, Kuester S, von Janda S. Socially (un)ac/ceptable errors of AI: consumer perceptions of different AI-induced errors. J Bus Res. 2025; 201: 115673.

[16] Rahman MA, Victoros E, Davis R, et al. Use of artificial intelligence in mental healthcare, health psychology, and related research: a narrative review to address challenges and opportunities. Health Sci Rep. 2025; 8: e71595.

[17] Khadangi A, Marxen H, Sartipi A, et al. When AI takes the couch: psychometric jailbreaks reveal internal conflict in frontier models. arXiv 2025; 2512.04124.

[18] Kéri S, Barkó B, Kelemen O. AI-derived Research Domain Criteria scores from medical records predict brain inflammatory markers in psychotic disorders: a cross-sectional, real-world study. Sci Prog. 2026; 109: 368504261417875.

[19] Kéri S, Kancsev A, Kelemen O. Algorithm-based modular psychotherapy alleviates brain inflammation in generalized anxiety disorder. Life (Basel) 2024; 14: 887.

[20] Kelemen O, Máttyássi A, Kéri Sz. Tailored treatment: opportunities for algorithm-based modular psychotherapy in Hungarian mental health care. [Egyénre szabva: az algoritmus-alapú moduláris pszichoterápia lehetőségei a hazai ellátásban.] Psychiatr Hung. 2025; 40: 172–183. [Hungarian]

[21] Adam D. Digital phenotyping using smartphones could help steer mental health treatment. Proc Natl Acad Sci USA 2025; 122: e2505700122.

[22] Martinez-Martin N, Greely HT, Cho MK. Ethical development of digital phenotyping tools for mental health applications: Delphi Study. JMIR Mhealth Uhealth 2021; 9: e27343.

[23] Gonzales B. App reads patient biometrics to detect mental health conditions. Available from: https://www.biometricupdate.com/202308/app-reads-patient-biometrics-to-detect-mental-health-conditions [accessed: 24 January, 2026].

[24] Garyfalli V, Kalisperakis E, Smyrnis A, et al. Smartwatch-derived digital phenotypes relate to psychopathology dimensions in patients with psychotic spectrum disorders: longitudinal observational study. JMIR Ment Health 2025; 12: e75774.

[25] Baydili İ, Tasci B, Tasci G. Artificial intelligence in psychiatry: a review of biological and behavioral data analyses. Diagnostics (Basel) 2025; 15: 434.

[26] Youyou W, Kosinski M, Stillwell D. Computer-based personality judgments are more accurate than those made by humans. Proc Natl Acad Sci USA 2015; 112: 1036–1040.

[27] Lan X, Han Z, Cheng Y, et al. Depression detection on social media with large language models. arXiv 2025; 2403.10750.

[28] Baran FD, Cetin M. AI-driven early diagnosis of specific mental disorders: a comprehensive study. Cogn Neurodyn. 2025; 19: 70.

[29] Birnbaum ML, Norel R, Van Meter A, et al. Identifying signals associated with psychiatric illness utilizing language and images posted to Facebook. NPJ Schizophr. 2020; 6: 38.

[30] Tong BG, Liang Z, He X, et al. AI-driven dynamic psychological measurement: correcting university student mental health scales using daily behavioral and cognitive data. Front Digit Health 2025; 7: 1615250.

[31] Okesanya OJ, Adebayo UO, Ngwoke I, et al. Artificial intelligence in psychiatry: transforming diagnosis, personalized care, and future directions. Explor Digit Heal Technol. 2025; 3: 101174.

[32] Funk M, Drew N, Cole C, et al. New WHO guidance on mental health and well-being across government sectors. World Psychiatry 2026; 25: 147–148.

[33] De Freitas J, Oguz-Uguralp Z, Kaan-Uguralp A. Emotional manipulation by AI companions. arXiv 2025; 2508.19258.

[34] Zhang Q, Zhang R, Xiong Y, et al. Generative AI mental health chatbots as therapeutic tools: systematic review and meta-analysis of their role in reducing mental health issues. J Med Internet Res. 2025; 27: e78238.

[35] Possati LM. Psychoanalyzing artificial intelligence: the case of Replika. AI Soc. 2023; 38: 1725–1738.

[36] Browne D, Slozberg M, Arthur M. Do mental health chatbots work? Available from: https://www.healthline.com/health/mental-health/chatbots-reviews [accessed: 24 January, 2026].

[37] Kneese T, Vecchione B, Marwick A. A chatbot for the soul: mental health care, privacy, and intimacy in AI-based conversational agents. Commun Change 2025; 1: 15.

[38] Cole-Turner R. Artificial intelligence and human spirituality: is a spiritual chatbot a good idea? Theology Sci. 2025; 23: 471–486.

[39] Fetrati H, Chan G, Orji R. Chatbots for sexual health improvement: a systematic review. Int J Hum Comput Interact. 2025; 41: 1997–2019.

[40] Yoon SC, An JH, Choi JS, et al. Digital psychiatry with chatbot: recent advances and limitations. Clin Psychopharmacol Neurosci. 2025; 23: 542–550.

[41] Harrer M, Miguel C, Tong L, et al. Effectiveness of digital interventions for eight mental disorders: a meta-analytic synthesis. Internet Interv. 2025; 41: 100860.

[42] Leung WK, Lam SC, Chan BC, et al. Chatbot interventions for improving mental health among people in Asia: a systematic review and meta-analysis of randomised controlled trials. BMJ Health Care Inform. 2026; 33: e101479.
[43] Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis. J Affect Disord. 2024; 356: 459–469.
[44] Li H, Zhang R, Lee YC, et al. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med. 2023; 6: 236.
[45] Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI. 2025; 2(4): DOI: 10.1056/AIoa2400802.
[46] American Psychological Association. Artificial intelligence and wellness apps alone cannot solve the mental health crisis. Available from: https://www.apa.org/news/press/releases/2025/11/ai-wellness-apps-mental-health [accessed: 24 January, 2026].
[47] Jamali L. ChatGPT shares data on how many users exhibit psychosis or suicidal thoughts. Available from: https://www.bbc.com/news/articles/c5yd90g0q43o [accessed: 24 January, 2026].
[48] Hudon A, Stip E. Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". JMIR Ment Health 2025; 12: e85799.
[49] Stokel-Walker C. AI driven psychosis and suicide are on the rise, but what happens if we turn the chatbots off? BMJ 2025; 391: r2239.
[50] Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017; 4(2): e19.
[51] Cuijpers P, Harrer M, Furukawa TA. Innovations to improve outcomes and uptake of psychotherapies for mental disorders: a state-of-the-art review. World Psychiatry 2026; 25: 4–33.
[52] McFadyen J, Habicht J, Dina LM, et al. Increasing engagement with cognitive-behavioral therapy (CBT) using generative AI: a randomized controlled trial (RCT). Commun Med (Lond). 2026 Jan 15. Doi: 10.1038/s43856-025-01321-8. Epub ahead of print.
[53] Avula VC, Amalakanti S. Artificial intelligence in psychiatry, present trends, and challenges: an updated review. Arch Ment Health 2024; 25: 85–90.
[54] Yirmiya K, Fonagy P. Mentalizing without a mind: psychotherapeutic potential of generative AI. J Med Internet Res. 2025; 27: e79156.
[55] Magee P, Ienca M, Farahany N. Beyond neural data: cognitive biometrics and mental privacy. Neuron 2024; 112: 3017–3028.
[56] Fedorczyk F. Expert comment: chatbot-driven sexual abuse? The Grok case is just the tip of the iceberg. Available from: https://www.ox.ac.uk/news/2026-01-14-expert-comment-chatbot-driven-sexual-abuse-grok-case-just-tip-iceberg [accessed: 24 January, 2026].
[57] Zimmerman JW, Ruiz AJ. Matters arising: a response to loneliness and suicide mitigation for students using GPT3-enabled chatbots. NPJ Ment Health Res. 2025; 4: 60.
[58] Boissin E, Costello TH, Spinoza-Martín D, et al. Dialogues with large language models reduce conspiracy beliefs even when the AI is perceived as human. PNAS Nexus 2025; 4: pgaf325.
[59] European Commission. AI Act (Shaping Europe's digital future) – Prohibited practices summary (Article 5). Available from: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai [accessed: 24 January, 2026].
[60] Perlis RH, Gunning FM, Usla A, et al. Generative AI use and depressive symptoms among US adults. JAMA Netw Open 2026; 9(1): e2554820. Erratum: JAMA Netw Open 2026; 9(2): e262242.

(Kéri Szabolcs dr.,
Sárospatak, Eötvös u. 7., 3950
e-mail: keri.szabolcs@unithe.hu)

„Lege artis medicinae."
(Az orvosi szakma szabályai szerint.)