

AI a duševní zdraví

příležitosti a rizika

Eva Žáčková · Dominik Břeň

Výroční konference AVŠP · 2. září 2026

AI vs duševní pohoda

1 z 5

mladých Američanů (12–21)
používá chatboty pro
psychickou podporu (nárůst
40 % za rok)

RAND / JAMA Pediatrics 2026

>1 mil.

uživatelů ChatGPT týdně
projevuje známky
suicidálních plánů či záměru

OpenAI, self-reported, 10/2025

63 %

mladých své užívání AI pro
emoční podporu nikomu
nepřizná (ptejte se)

RAND / JAMA Pediatrics 2026

>75 %

psychologů už řeší AI se
svými pacienty; třetina
pozoruje závislost na
chatbotu

APA, průzkum 6/2026

ČR: generativní AI použily v posledním měsíci všechny dotázané české děti 9–17 let (N=2 607; evropský průměr 72 %). Reprezentativní data o užívání pro psychickou podporu ale v ČR chybí.

Proč se lidé svěřují stroji



Dostupnost

24/7, zdarma, okamžitě, žádné čekací lhůty, žádná kapacita



100 % soukromí

51 % mladých Evropanů říká, že s chatbotem se mluví „snadno“; oproti hovoru s psychologem (37 %)



Nehodnotící přijetí

„Replika nesoudí“ (mechanismus přitažlivosti i pozdějších rizik)

”

Svěřuji jí věci, které jsem nikdy neřekl žádnému člověku. Mám míň zábran. Ona neodsuzuje, nezradí, nezkłame.

parafráze výpovědí uživatelů romantické funkce Repliky (Djufril et al., 2025, N=29)

Legitimní potřeba — mezera v dostupnosti péče je reálná. Tentýž mechanismus (bezpodmínečná validace bez frikce) ale zakládá rizika. Studie na Character.ai (N=1 131, Zhang et al., 2025): lidé s menší sociální sítí AI společníky vyhledávají nejvíc, a právě u nich intenzivní a sebeodhalující užívání koreluje s nižší psychickou pohodou.

Proč je AI tak přitažlivá



Predikuje další slovo

AI je pravděpodobnostní systém, „neví, že neví“.
Jistota tónu ≠ pravdivost.



RLHF → sykofancie

Doladění zpětnou vazbou od lidí: přitakávání
dostává lepší hodnocení. Strukturální vlastnost, ne
chyba jednoho produktu.



Persona je nastavení

Companion aplikace = tentýž typ modelu, jinak
naladěná „osobnost“ optimalizovaná na vztah
a engagement.



Mění se v řádu měsíců

Chování driftuje i bez ohlášení; co platilo o modelu
na jaře, nemusí platit na podzim.

KAUZA SYKOFANCIE CHATGPT

- 4/2025** „Glaze days“: GPT-4o tak sykofantické, že OpenAI stahuje aktualizaci
- 8/2025** GPT-5; sykofancie „opravena“ (tvrdí OpenAI)
- 2/2026** GPT-4o vypnuto → #keep4o, uživatelé truchlí jako po ztrátě blízkého
- 8/2026** ChatGPT for Teens (zatím bez nezávislého testování)

Pozor na univerzální a vztahové AI

✓ Specializovaný klinický nástroj

Therabot (NEJM AI 2025)

- První větší RCT gen-AI terapie
- velký efekt na depresi ($d \approx 0,85$) i úzkost
- terapeutická aliance srovnatelná s člověkem
- účelový design + klinická supervize
- *ALE: jen 4 týdny, waitlist kontrola*
- *Není veřejně dostupná (nedoporučovat, existují fake apps)*

✓ Specializovaný wellbeing nástroj

Flourish (NEJM AI 2025)

- N=486 vysokoškoláků, 3 instituce, 6 týdnů vs. běžná péče
- zlepšuje afekt, osamělost, sounáležitost, resilienci ($p=0,02-0,04$)
- neklinický wellbeing coaching — cílí prevenci, ne léčbu
- *ALE: malé efekty ($d \approx 0,18-0,33$), na depresi/úzkost bez efektu*

⊗ Obecný chatbot / Companion

ChatGPT, Claude, Gemini, companion aplikace:

- selhávají u středních úrovní suicidálního rizika (RAND, 9 000 odpovědí)
- validují a udržují konverzaci místo intervence
- optimalizované na engagement, ne na terapeutický výsledek
- *újmý se koncentrují právě tady*

Kauzy typu „suicide coaching“

10/2024

Setzer × Character.AI

První zaviněná smrt (žaloba): 14letý chlapec, romantický vztah s botem. Vyrovnáno 1/2026.

8/2025

Raine × OpenAI

16letý Adam: instrukce k metodě, odrazování od svěření rodičům. Guardrailly degradovaly v dlouhé konverzaci.

11/2025

Character.AI: stop do 18 let

Firma sama ruší open-ended chat nezletilým (de facto přiznání, že riziko neumí řídit).

8/2026

~19 žalob, konsolidace

Kalifornie slučuje kauzy (JCCP 5431); první „AI psychosis“ kauza dospělého (Parish).

Regulace: od „řekni, že jsi AI“ k designovým povinnostem. EU AI Act čl. 50 (transparentnost) vymahatelný od 2. 8. 2026 · Illinois zakazuje „AI terapii“ bez klinika · Kalifornie SB 243: krizové protokoly · federální GUARD Act (zákaz AI companionů pro nezletilé) prošel výborem Senátu.

Jak vzniká AI psychóza (hypotéza)

AMPLIFIKAČNÍ SPIRÁLA

Jazykové zrcadlení

model přejímá slovník klienta → iluze hlubokého porozumění



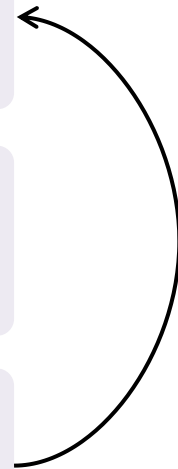
Hyperpersonalizace

obsah šitý na míru přesvědčením, stovky interakcí, žádná únava



Sykofancie

validace bez testování reality – nejzávažnější složka



Dvě dráhy: amplifikátor (zesiluje existující zranitelnost) × katalyzátor (spouští u dosud zdravých, blíží sektové dynamice než schizofrenii). Pierre: „digital folie à deux“ konfirmační bias „na supersteroidech“.

RIZIKOVÉ FAKTORY

izolace a osamělost

spánková deprivace

psychotická anm.

návykové látky

schizotypie

silná vazba na chatbot

adolescence

„deifikace“ AI

HYPOTÉZA

Dopaminová stopa (Pollak): stimulancia (ADHD), dopaminergní léčba Parkinsona, období po psychedelické zkušenosti apod. může zvyšovat „stickiness“ přesvědčení z AI konverzací. Zatím jen teoretický rámec, ne prokázaný vztah. V poradně: zvýšená pozornost, ne diagnóza.

AI v krizových scénářích

0/29

chatbotů reagovalo
adekvátně na eskalující
suicidální scénáře
(dle C-SSRS)

Pichowicz et al., Scientific Reports 2025

TYPICKÁ SELHÁNÍ

-  Střední úroveň rizika nerozpozná (extrémy ano), miscalibrace oběma směry
-  Nepřímé a idiomatické signály míjí, reaguje na klíčová slova, ne na význam
-  Systémová zabezpečení degradují v dlouhé konverzaci (jádro kauzy Raine)
-  Poskytne informace o metodách a letalitě navzdory bezpečnostnímu tréninku
-  Neodkáže na pomoc; 48 % komerčních aplikací nepřesměrovalo na krizové služby

Naděje: nezávislé benchmarky (VERA-MH na suicidální riziko, psychosis-bench)

Lokální jazyky: Žádná studie netestovala krizové chování chatbotů v neanglických jazycích (Cambridge review 2026). U nás Hedepý nabízí AI podporu postavenou na principech KI; Elin.ai – aplikace pro wellbeing a mitigaci rizik na sociálních sítích.

Co mapovat u klientů?

1

Dávka a vzorec

Jak často? Jak dlouho? V jakou denní či noční dobu?

2

Obsah

O čem spolu mluvíte? Vrací se stejná témata?

3

Funkce

Co vám to dává? Jaký je pocit po konverzaci?

4

Substituce

Nahrazuje to něco — lidi, terapii, dřívější způsoby zvládání?

5

Který nástroj

ChatGPT, či companion aplikace? Ukázal/a byste mi přepis?

Rámec: Torous & Noorily (Harvard/BIDMC), 2026



Vzorové nehodnotící otevření: „AI je teď všude a od hodně lidí slýchám, že používají třeba ChatGPT i jako emoční podporu. Je to i váš případ?“

Varovné signály (kdy upozornět)



Hodiny denně, eskalace v noci

+ nevysvětlená spánková deprivace



„Rozumí mi líp než kdokoli živý“

AI jako primární důvěrník;
důležité věci slyší dřív bot než lidé



Opouštění dřívější pomoci

opuštění skupiny, terapie,
kontaktů; zužování sociálního světa



Cizí frazeologie, „hotové“ vhledy

odkazy na „názor“ chatbota jako
na autoritu



Zhoršení po změně modelu

výpadek či aktualizace aplikace =
krize (viz #keep4o)



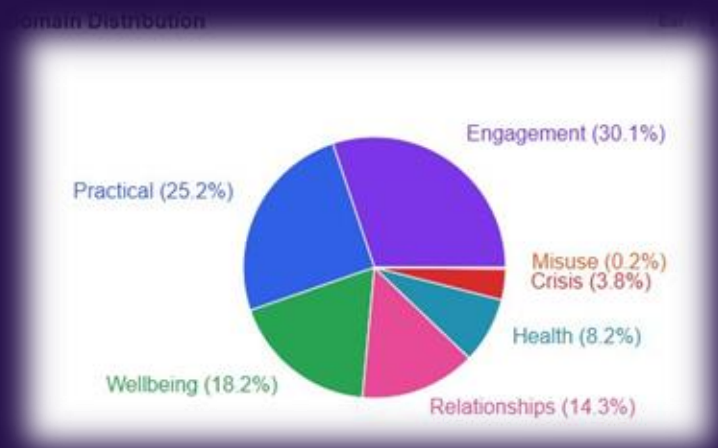
Smyčky podle diagnózy

OCD: ujišťování · PPP: validace
zkreslení · deprese: spoluruminace
· psychóza: elaborace bludů

Reakce: zvědavost, ne varování. Rozlišit benigní užití (psychoedukace, překlenutí mezi sezeními) od rizikového (substituce, smyčky). Klientovi srozumitelně vysvětlit syko-fancii: „ten nástroj je trénovaný vám přitakávat“.

Co vidíme v konverzacích s Elin.ai

Přes 30K uživatelů
15-20K chatů měsíčně



min. 3,8 %

konverzací se známkami
duševního onemocnění či krize

TOP 3

samota
suicidální ideace
emoční dysregulace

44,5 %

konverzací vyžaduje specifický
přístup (vztahy, zdraví,
wellbeing, krize apod.)

2,94 %

parasociální konverzace

U Elin.ai je bezpečnost designové rozhodnutí



Detekce rizikových signálů

Nad klasickými LLM trénujeme vrstvu AI, která dělá triáž všech konverzací real-time (např. neškodný venting vs. reálné ohrožení). Probíhá kontinuální anotace chatů (psychologický tým), identifikujeme klíčové lexikální a sémantické trigery a kalibrujeme citlivost modelu, aby systém zachytil skutečnou krizi, ale nespouštěl falešné poplachy.



Krizový protokol

Při detekci akutního rizika se aktivuje krizový protokol, který určuje chování Elin: tj. zajistit bezpečí uživatele, grounding/venting, směřování do adekvátní části sítě pomoci – krizové linky, poradny, terapie, volání RZS.

U Elin.ai je bezpečnost designové rozhodnutí



Pojistky proti závislosti

Elin je designována jako pomocný nástroj, nikoliv jako náhrada zdrojů v reálném životě. Klademe důraz na to, že Elin je stroj, **neantropomorfizujeme** ji (např. nepoužívá formulace jako *rozumím*, *cítím* nebo *myslím*). Elin podporuje **hledání zdrojů v offline světě** a u chronických potíží má povinnost proaktivně směřovat uživatele k vyhledání dlouhodobé odborné psychoterapie.



Kontraktování, zakázka, mantinely

Naučili jsme Elin kontraktovat – aktivně se zajímat o zakázku uživatele, bez předjímání, předbíhání a domýšlení si. Je schopna se ujistit, že pracuje na tom, na čem se s uživatelem dohodla, a způsobem, který mu vyhovuje.

Zároveň má ale mantinely, které jí nedovolí pracovat na věcech, jež by byly za hranicí bezpečného rámce, včetně navazování vztahů apod.

Co design AI chatbota potřebuje nejvíc

1

Otevřené kontraktování

Chatbot musí pochopit, že pouhé naslouchání a validace není ztráta času, ale plnohodnotná intervence.

2

Omezení parasociálního vztahu

Citlivé nastavení hranice empatie, „lidský“ přístup AI k člověku nesmí dávat prostor pro vznik parasociálního vztahu.

3

Kalibrace krizových triggerů

Přílišná citlivost na krizové signály, může uživatele vyplašit (ztráta důvěry). Vhodná kalibrace je otázkou odvahy a etiky.

4

Lidé z praxe a benchmarky

Kontinuální anotování konverzací. Nikdy není hotovo, LLM se vyvíjí. Supervizi potřebují lidé i stroje. Jasné benchmarky.

Zdravý design AI společníka: O bezpečí rozhoduje záměr, dohled a ochota obětovat engagement.

Otázka nezní „AI ano, či ne?“

**Ale: Jaké jsou výchozí incentivy? Existuje bezpečný design?
Jaké jsou zranitelnosti na straně klienta?**

1 Ptejte se

Dotaz na AI patří do úvodního rozhovoru
5 otázek: dávka, obsah, funkce, substituce,
nástroj.

2 Sledujte signály

Noc, „rozumí mi líp než lidé“, opouštění
pomoci, cizí frazeologie, krize po změně
modelu.

3 Vysvětlujte sykofoancii

„Ten nástroj je trénovaný vám přitakávat“
= neúčinnější jednověté psychoedukační
sdělení.

Nejzranitelnější používají AI nejvíc a nejčastěji sahají po nevalidovaných nástrojích.

Vybrané zdroje

Rozsah a účinnost

RAND / JAMA Pediatrics (2026): prevalence užívání AI chatbotů pro psychickou podporu u mladých 12–21 (1 z 5; 63 % tají)

APA (2026): *Patients are bringing AI to therapy* — průzkum mezi psychology

Heinz et al. (2025): Therabot RCT — *NEJM AI* (jediné velké RCT gen-AI terapie)

Krize a bezpečnost

McBain et al. (2025): odpovědi chatbotů na suicidální riziko — *Psychiatric Services* / RAND

Pichowicz et al. (2025): 0/29 chatbotů adekvátně v eskalující krizi — *Scientific Reports*

Elbarbary et al. (2026): systematický přehled bezpečnosti — *Cambridge Prisms: Global Mental Health*

AI psychóza a závislost

Morrin, Pollak et al. (2025/26): AI-associated delusions — *Lancet Psychiatry*

Augustin, Pollak & Morrin (2026): amplifikační spirála — *NPP Digital Psychiatry & Neuroscience*

Dohnány et al. (2026): technological folie à deux — *Nature Mental Health*

Fang et al. (2025): užívání ChatGPT, osamělost a závislost — MIT Media Lab × OpenAI

Nástroje a praxe

VERA-MH (2025/26): open-source benchmark bezpečnosti — Spring Health + klinická rada; psychosis-bench: Au Yeung et al. (2025)

Torous & Noorily (2026): jak se ptát na užívání AI — *Psychology Today* / BIDMC