Designing for Long-term Impact: A Hybrid Intelligence Case Study in AI-Assisted Mentalizing-Based Therapy

Jacob SHERSON[a,1], Janet RAFNER[b], Sune Bo[c], Bianka SZÖLLÖSI[a], Florent VINCHON[e], Paul-Arno LAMARQUE[d], Florian DE BONI[d]
[a] *Center for Hybrid Intelligence, Aarhus University*
[b] *University of Southern Denmark*
[c] *Psychiatric Research Unit, Region of Southern Denmark & University of Copenhagen*
[d] *Hybrid-iQ*
[e] *Lab. de Psych. et d'Erg. Appl. (LaPEA), Univ. Paris Cité and Univ Gustave Eiffel*

Abstract. AI deployment in customer-facing service sectors increasingly bypasses domain experts in favor of direct automation, extracting tacit professional knowledge without returning value to the profession. We propose a two-layer Hybrid Intelligence design framework built around two mutually constitutive questions: how will building this tool be good for the domain expert, the profession, and the world; and how do we build an interaction in which the domain expert learns more from the system than the system learns from the expert. The long-term vision grounds design by building psychological safety and co-creation willingness; concrete interface elements in turn inspire and concretize that vision. We illustrate the framework through MentaBot, an AI chatbot co-developed with a clinical psychologist for Mentalization-Based Therapy, showing that a process grounded in long-term professional and planetary vision, structured expert co-creation, and metacognitive interface design produces an interface that meaningfully outperforms generic large language models — not because the underlying model is more powerful, but because the human expertise shaping it was more fully brought to bear.

Keywords. Mentalizing, Therapy, GenAI, Hybrid Intelligence, chatbot

1. Introduction

The willingness to co-create with AI systems is a behavioral disposition that must be actively cultivated [1]. Psychological safety — feeling one's expertise is developed rather than extracted — is essential for sustained expert engagement in genuine co-creation [2]. The narrative framing around an AI system shapes user interaction and learning patterns [3,4]. These findings highlight a gap between AI-first automation and expert-centered AI, but no design logic has integrated them. A recent case illustrates this gap: when Verizon enhanced AI-assisted drafts with human agent touch during a January 2026 outage, satisfaction rose to 88% versus 60% for AI-only interactions [5]. Human presence, structurally embedded in the interface, produced better outcomes at scale - yet this design logic remains unarticulated in most processes.

[1] Corresponding Author: Jacob Sherson, sherson@mgmt.au.dk

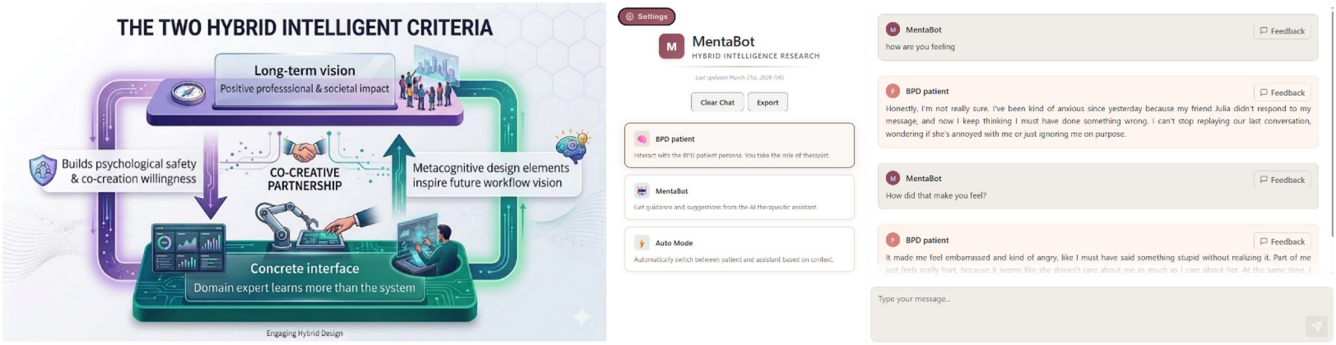

Figure 1. Left: the two mutually reinforcing design criteria of a hybrid intelligent interface — positive long-term professional and societal impact, and an interaction in which the domain expert learns more than the system. Right: the MentaBot interface with three interaction modes — therapist dialogue, BPD patient persona, and autonomous agent-to-agent conversation — and configurable system parameters.

Drawing on the Hybrid Intelligence (HI) manifesto [6], a consensus definition across more than 40 interdisciplinary researchers, which defines HI as "a co-created, deeply interactive human-AI system that empowers humans to deliberately improve workflows within a hybrid learning organization pursuing competitiveness within a human-touch marketplace", here, we propose two practitioner-oriented design questions (Fig.1 Left):

- *DQ1: how will building this tool be good for the domain expert, the profession, and the world?*
- *DQ2: how do we build an interaction in which the domain expert demonstrably learns more from the system than the system learns from the expert?*

2. Long-term professional and planetary grounding

Borderline Personality Disorder (BPD) affects 1.6% of the general population, with mortality ten times higher than the background population and 80–95% of BPD patients engaging in self-harm [7, 8]. Mentalization-Based Treatment (MBT), which aids understanding behavior and actions through mental states, shows moderate but persistent reduction of self-harm [9]. The critical issue for the patients is the gap between therapy sessions, when crisis escalation often occurs. MentaBot is an MBT-based AI chatbot co-developed with clinical psychologists to support BPD patients during high-risk intervals.

Addressing DQ1 convincingly requires concrete future workflow design, ideally modeling economic outcomes of hybrid versus automated pathways and securing leadership commitment to make the HI future real. Concretely, MentaBot is designed as a training tool to dramatically speed up therapeutic expertise accumulation but primarily a bookable between-session therapeutic offering. The therapist retains case formulation, crisis judgment, and the therapeutic alliance that is itself a documented driver of MBT outcomes [10]. The long-term professional vision goes considerably further, however. The metacognitive design elements described in the next section open the possibility of an entirely revisioned in-clinic workflow [11]. Rather than serving the patient alone between sessions, these dynamically generated dimension indicators could be displayed discretely to the therapist (e,g, via AR glasses) providing in-session guidance to scaffold the therapist's clinical judgment. More transformatively, they could be displayed to both

patient and therapist during live sessions, creating a digitally mediated shared therapeutic space. Mentabot furthermore contains a complementary value pathway: Mentalizing is a fundamental human skill linked to increased social functioning and improved mental health [12] and broad application in schools would offers substantial societal benefits [13]. This places MentaBot in a rare category: one of the few AI cases where even a worst-case automation scenario — reducing demand for MBT therapists — might still produce a net global positive. The best case, benefiting both the clinical profession and society, is the design target. This alignment, from individual outcomes to planetary impact, generates the lead clinician's enthusiasm for the co-creation work described next.

3. The interface: co-creation process, metacognitive design and evaluation

The co-creation process: The long-term professional and planetary alignment generates psychological safety and co-creation willingness [1,2], propelling the lead clinician from a consultant, to a design authority with genuine decision power over the therapeutic philosophy the system must embody. Concretely, rather than providing open-ended feedback, the clinician and his colleagues evaluate each MentaBot iteration against a structured taxonomy of named failure modes. For each version the evaluator assesses which failure modes are resolved, which remain, and critically whether fixes in one area have introduced unintended degradation elsewhere — a regression-aware review that transforms tacit clinical expertise into an explicit, repeatable evaluation protocol that itself constitutes professional learning.

Metacognitive design elements (DQ2): MentaBot's central interface commitment is to make non-mentalizing modi (pretend mode, psychic equivalence, and teleological thinking) as well as the four constitutive dimensions of mentalizing [14] (self/other, cognitive/affective, internal/external, and implicit/explicit) dynamically visible during the interaction, prompting the patient to reflect on their own mentalizing state rather than simply receiving a therapeutically framed response. This is realized through a multi-agentic design with each agent specialised for one task, e.g. analysing one precise aspect during the conversation. The patient-facing design is thus not only a clinical intervention but a stepping stone toward the future hybrid intelligent professional workflow.

Prototype evaluation: MentaBot vs ChatGPT: Structured expert evaluation using identical patient statements reveals a clear performance gap between MentaBot and ChatGPT (GPT-5.2, April 2026), assessed against the failure mode taxonomy developed above. When a patient says "I feel awful these days" and subsequently reports anger, MentaBot stays with the presented feeling through single focused curious questions. ChatGPT categorizes the feeling, offers breathing techniques, and asks multiple questions simultaneously — solving rather than scaffolding the capacity to mentalize. When a patient describes simultaneously wanting to hit and hug his girlfriend after an argument, MentaBot asks one precisely targeted question about that ambivalence, and further prompts mentalizing in the patient, whereas ChatGPT again responds with analytical sophistication. The expert verdict: MentaBot stays curious and effectively focused, scaffolding the patient's own mentalizing capacities so crucial for adequate social functioning [15]; ChatGPT loses focus and intervenes where it should inquire. These contrasts are the direct output of the co-creation process — the failure mode taxonomy makes the gap legible, and the framework predicts it: long-term vision, genuine co-creation willingness, and metacognitive design commitments produce an interface that AI-first development cannot replicate.

References

[1] Mao Y, Rafner J, Wang Y and Sherson J (2024) HI-TAM, a hybrid intelligence framework for training and adoption of generative design assistants. Front. Comput. Sci. 6:1460381. doi: 10.3389/fcomp.2024.1460381, https://doi.org/10.3389/fcomp.2024.1460381

[2] Sherson, J., Rafner, J. F., Holm, A. B., & Marler, J. (2025). A 4P hybrid intelligent development and organisational change framework. In HHAI 2025, IOS Press. https://ebooks.iospress.nl/doi/10.3233/FAIA250687

[3] Rafner, J., Guloy, R. Q., Wen, E. W., Chiodo, C. M., & Sherson, J. (2025). From interaction to collaboration: How hybrid intelligence enhances chatbot feedback. arXiv preprint arXiv:2504.13848.

[4] Rafner, J., Tudoran, A. A., Beliatis, M., & Sherson, J. (2025). Framing the future of work: How narrative influences perceptions of AI. In HHAI 2025 (pp. 528-530). IOS Press. https://ebooks.iospress.nl/doi/10.3233/FAIA250683

[5] Bariso, J. (2026, January 15). After Verizon went down, customers were furious. Verizon's response is a lesson for every business. Inc. https://www.inc.com/justin-bariso/verizon-down-customers-furious-response-lesson-every-business-emotional-intelligence/91288660

[6] Sherson, J, et al, The Hybrid Intelligence Manifesto: A Multi-Level Socio-Technical Framework for Keeping Human Judgement Economically Indispensable, in submission, online version available at: https://hi-infinity.ai/manifesto/

[7] Fok ML, Hayes RD, Chang CK, Stewart R, Callard FJ, Moran P. Life expectancy at birth and all-cause mortality among people with personality disorder. J Psychosom Res. 2012 Aug;73(2):104-7. doi: 10.1016/j.jpsychores.2012.05.001. Epub 2012 May 26. PMID: 22789412.

[8] Goodman M, Tomas IA, Temes CM, Fitzmaurice GM, Aguirre BA, Zanarini MC. Suicide attempts and self-injurious behaviours in adolescent and adult patients with borderline personality disorder. Personal Ment Health. 2017 Aug;11(3):157-163. doi: 10.1002/pmh.1375. Epub 2017 May 22. PMID: 28544496; PMCID: PMC5571736.

[9] Hajek Gross C, Oehlke SM, Prillinger K, Goreis A, Plener PL, Kothgassner OD. Efficacy of mentalization-based therapy in treating self-harm: A systematic review and meta-analysis. Suicide Life Threat Behav. 2024 Apr;54(2):317-337. doi: 10.1111/sltb.13044. Epub 2024 Jan 27. PMID: 38279664.

[10] Bateman, A., Campbell, C., Luyten, P., & Fonagy, P. (2018). A mentalization-based approach to common factors in the treatment of borderline personality disorder. Curr Opin Psychol, 21, 44-49. https://doi.org/10.1016/j.copsyc.2017.09.005

[11] See further details and graphical illustrations of the envisioned future professional pathways of the Mentabot interface as well as updated details on the Mentabot scientific project at https://hi-infinity.ai/projects/mentabot/

[12] Fonagy, P., Campbell, C., & Luyten, P. (2023). Attachment, Mentalizing and Trauma: Then (1992) and Now (2022). Brain Sciences, 13(3). https://doi.org/10.3390/brainsci13030459

[13] Valle, A., Massaro, D., Castelli, I., Sangiuliano Intra, F., Lombardi, E., Bracaglia, E., & Marchetti, A. (2016). Promoting Mentalizing in Pupils by Acting on Teachers: Preliminary Italian Evidence of the "Thought in Mind" Project. Front Psychol, 7, 1213. https://doi.org/10.3389/fpsyg.2016.01213

[14] Bo, S., Sharp, C., & Lind, M. (2025). Maladpative social cognition and perspective-taking. In: ICD-11 personality Disorder. Ed. Bach, B. Oxford University press. pp 213-224)

[15] Luyten, P., Campbell, C., Allison, E., & Fonagy, P. (2020). The Mentalizing Approach to Psychopathology: State of the Art and Future Directions. Annu Rev Clin Psychol, 16, 297-325. https://doi.org/10.1146/annurev-clinpsy-071919-015355