Establishing ethical design principles for 'Humanised' generative AI workplace coaches

Jonathan Passmore, Bergsveinn Olafsson, Jazz Rasool & Tara Wilson

Citation: Passmore, J., Olafsson, B., Rasool, J. & Wilson, T. (2025). Establishing ethical design principles for 'Humanized' generative AI bots in workplace coaching conversations. *The Coaching Psychologist,* 21(1), 48-58. https://doi.org/10.53841/bpstcp.2025.21.1.48

Abstract

The purpose of this conceptual paper is to explore the design principles which can be used to inform the design and deployment of AI coaching agents, an area of growing interest and activity. The paper draws on coaching psychology, AI ethics and empirical studies of AI coaching to offer seven design principles: (i) transparency and informed continuous consent, (ii) autonomy and user empowerment, (iii) privacy and data protection, (iv) fairness, (v) do no harm and beneficence, (vi) accountability and (vii) alignment. Further it argues that generative AI tools should be viewed not as agency free but as more akin to a domesticated pet and that liability for unintended negative events, such as harm to a person, should be shared between the designers, the IP holders and the AI agent. The paper aims to inform this emerging agenda for designers, organisational buyers and coaching psychologists who are considering or deploying AI workplace coaching agents.

Keywords AI agents, AI Coaching

Introduction

OpenAI's release of a voice-driven GPT tool introduced a mainstream debate around machine humanisation and intellectual property rights (New York Times, 2024). Prior instances, including BIPOC Nazi soldiers (New York Times, 2023), suicide collusion (Brussels Times, 2023), and deceptive outputs (Greenblatt et al., 2024), have raised urgent questions about the ethics of generative AI design. Among these emerging technologies is the development of AI coaching agents (coachbots), which aim to replicate human coaching dialogues and encourage clients to reflect on workplace challenges, develop new insights, learn and identify future actions. While scholars note definitions of coaching continue to evolve (Passmore & Lai, 2019), the International Coaching Federation (2020) defines coaching as a partnership which fosters insight and potential. However, we have argued that such definitions do not fully capture the intimate, confidential and high-trust nature of coaching conversations (Passmore & Abri von Bartheld, 2024).

Some authors suggest AI might eventually replace human coaches for standardised topics (Rauen, 2018), while others propose AI might assist or augment coaching (Greif, 2018; Oesch, 2018; Passmore & Tee, 2023; Smith et al., 2025). A third group rejects the possibility of machine-led coaching altogether, maintaining that coaching is inherently relational (Bachkirova & Kemp, 2024). Evidence suggests the quality of these tools is developing fast, with some tools assessed as meeting the accreditation standards set by professional bodies (Passmore, Tee & Rutschmann, 2025).

As AI development continues, professional human coaches have expressed anxiety, viewing AI as an existential threat (Diller, Stenzel & Passmore, 2024). The rapid pace of change has seen AI tools entering the agenda for learning and development, and we predict that by the end of the 2020s most large organisations will be using generative AI tools as an integrated part of their talent development pathways. There is thus an urgent need to inform a nd set standards for developers, buyers and users. In this conceptual paper we draw on a narrative review of coaching psychology, AI ethics frameworks and empirical studies of AI coaching to articulate practically oriented design principles for generative AI coaching agents (Floridi & Cowls, 2019; Passmore et al., 2025; Passmore & Tee, 2023).

Defining AI Coachbots

The terminology surrounding AI in coaching remains contested (Graßmann & Schermuly, 2021). Diller and Passmore (2023) offered the following definition: "*a synchronous coaching experience, where the machine replaces the role of the human coach, facilitating their human client in goal setting, issue exploration, personal reflection and developing insights and actions*". In this paper we focus on AI workplace coaches, by which we mean agents that support employees with development goals, performance challenges and career reflections rather than broader mental health or therapeutic concerns (Passmore & Tee, 2023).

Evaluating Impact

Empirical evaluations of AI in coaching and broader workplace learning remain limited. General tools like ChatGPT display variable performance, making it difficult to anticipate areas where AI will enhance productivity or deliver high-quality outputs (Dell'Acqua et al., 2023). A systematic literature review of AI in coaching identified only sixteen studies, mainly within education and healthcare settings and often using student populations (Passmore, Olafsson, & Tee, 2025) and often using student populations. Across studies in wellbeing, physical activity, anxiety and weight management, conversational agents have shown modest but positive effects on outcomes such as resilience, physical activity, exam anxiety and weight management (Ellis-Brush, 2021; Figueroa et al., 2021; Hassoon et al., 2021; Kannampallil et al., 2022; Mai et al., 2022; Stephens et al., 2019). During the Covid-19 lockdown, Terblanche et al. (2022) demonstrated comparable goal achievement outcomes between AI bots and human coaches, but given that the AI coach data was collected during a Covid-19 lockdown, care must be taken when seeking to generalise from these results. More recently, papers have used professional human coach assessment frameworks to evaluate AI coach capabilities and found AI coaching tools are able to pass these assessments and secure professional coach accreditation (Passmore, Tee & Rutschmann, 2025). In summary, results from early applications are promising, but ethical concerns about generative AI tools limit their adoption by organisations (Passmore & Tee, 2023).

Design Principles: Generation 1 (Coachbots) to Generation 3 (AI Coaching agents)

Design Principles: Generation 1 (Coachbots) to Generation 3 (AI Coaching agents)

Early AI coachbots (Generation 1) were largely text based, driven by a script and most adopted a machine identity, through their designer's choice of name. At this stage of development, researchers proposed four design principles for AI bots (i) outcome efficacy (ii) use of recognised theoretical models (iii) ethical conduct and (iv) narrow coaching focus (Terblanche, 2020). The framework has provided a useful starting point, and has informed both research studies, (Terblanche et al., 2022; Terblanche & Cillers, 2020) and exploration of ethical challenges in workplace AI coaching (Passmore and Tee, 2023).

Emerging technologies – Generation-2 and Generation-3 'AI Agents' - Post 2023

Since their emergence in the late 2010's AI technologies have continued to develop, and what was experimental work by small start-ups has become deployed of commercial products by 2024: Generation-2 AI Coaching bots. Similar AI tools are appearing in a range of other service industries from holiday bookings to call centres. These offer some or all of the following features: use a human name, have a static 2D human image, and use a human voice for interactions.

The period 2025-2030 will see a coming together of technologies, in what might be considered to be third generation: AI Coaching Agents. As AI technologies develop towards Artificial General Intelligence (AGI) (Feng, et al., 2024) new challenges emerge. These newly emerging agents could combine AI, with VR technologies such as avatars and voice replication, to create a moving image, that feels to the user as if it's a real person. The agent may project the persona or image of a celebrity, or respected coaches, be available in moving 2D (see: Synthesia, 2024) or 3D using photorealistic codec (see: Fridman, 2023) and offering users the option to be coached by a global leadership guru or a film celebrity of their choice, or allow users to select their choice of coach based on gender, ethnicity, language or other aspects of their identity.

The introduction of AI coaches that are voice driven and combined with photorealistic human image avatars of the user's choice will represent a significant shift in the user experience, changing how the user perceives the bot and engages with it. The emerging anthropomorphism of AI coaching raises new ethical concerns (Kaur & Gupta, 2024; Yu et al., 2025). We believe the emergence of Generation-3 AI Coaching Agents brings urgent consideration of the ethical principles which designers, individual users and organisations should consider.

Ethical design principles for humanised AI coaching agents

Seven ethical design principles (Table 1) are proposed to inform the design and deployment of humanised AI coaching agents: transparency and informed consent, autonomy and user empowerment, privacy and data protection, fairness, beneficence, accountability and alignment.

Table1: Design Principles – Third generation AI Coaching Agents

Principle 1: Transparency and informed, continuous consent
Principle 2: Autonomy and user empowerment
Principle 3: Privacy and data protection
Principle 4: Fairness
Principle 5: Do no harm and Beneficence
Principle 6: Accountability
Principle 7: Alignment

Principle 1: Transparency and Informed, Continuous Consent

Transparency and informed, continuous consent underpin all high-trust interactions, such as coaching, mentoring and counselling. Users should receive clear, prior disclosure of an AI coaching agent’s nature, capabilities and constraints, including its role, the data it gathers, modes of processing, retention period and intended purposes (Turilli & Floridi, 2009). As generative models “hallucinate”, producing inaccurate or misleading responses (Alkaissi & McFarlane, 2023), developers should alert clients to the risk of such errors, flagging the possibility of false information with each output (Lovejoy, 2019).

Table 2: Transparency Assessment questions

To what extent does the tool:

I. disclose it is an AI's tool, its role, what data it collects, how it processes the information, how long data will be held and the purposes for which it will be used?

II. disclose the tool may make errors, such as false assumptions, false claims or provide misleading information which should be checked by the user

III. explicitly seek and refresh informed consent before proceeding based on a summary of the key points and easy access to plain language terms and conditions

IV. provide information for users in plain language on design principles such as bot decision making

V. offer users validation through external accreditation

Genuine transparency demands explicit, straightforward informed consent rather than implicit acceptance buried in lengthy terms and conditions. Consent should be obtained at the point of access through concise explanations stating how data will be stored, who may access it and when deletion will occur, thus enabling an informed decision rather than a superficial "cookie" acknowledgement. Developers should therefore provide intelligible summaries of system functions, limitations and privacy provisions before and during use (see Table 2).

Principle 2: Autonomy and User Empowerment

Autonomy is a fundamental psychological requirement and a source of intrinsic motivation and well-being (Ryan & Deci, 2000). Ethical AI for high-trust conversations must therefore preserve, rather than erode, user self-determination. An AI coaching agent should strengthen the client's capacity to choose, not steer them towards predetermined solutions or foster dependence. Early and continuous involvement of users and coaching experts in design and testing helps align systems with client expectations and mitigates the exploitative patterns seen on social media, which have been linked to loneliness and psychological distress (Hunt et al., 2018).

Supporting autonomy demands several design choices. AI coaching agents ought to refrain from prescribing advice, recognising that individual differences in personality, environment and culture mean that no single course of action fits every circumstance. The agent should emulate an experienced human coach by cultivating critical reflection, posing questions and clarifying consequences, while remaining firmly non-directive. Should a system venture into advisory territory, it should be recast as a mentor or training tool so that its prescriptive role is explicit.

Developers must additionally furnish clear educational material explaining the agent's scope, decision processes and optimal use. Independent audits and professional kitemarks can provide impartial evaluations, assisting users who lack the expertise or time to assess such technologies themselves (see Table 3).

Table 3: Autonomy Assessment questions

To what extent does the tool:

I. Encourage user autonomy

II. Have users been involved in design and piloting

III. Avoid providing advice

IV. Stay within the role of a coach

Principle 3: Privacy and Data Protection

The use of AI coaching agents involves the collection and processing of sensitive personal data. As AI agents become more powerful, there are increasing economic incentives on AI developers to collect data that can invade personal privacy (Dwivedi et al., 2023). This challenge is one already experienced by users of social media. Protecting the privacy of users is paramount to maintaining ethical standards. AI systems should adhere to data protection policies, ensuring that user data is collected, stored and processed in a secure and confidential manner (Ryan, 2020). This will become increasingly important if users are to trust the AI systems in conversations where trust is a prerequisite.

Privacy concerns extend beyond data security to include the ethical use of collected data. Users should be informed about how their data will be used and given the option to opt-out of data collection for purposes unrelated to the primary function of the AI system.

The ethical handling of data also involves ensuring that users have control over their own data, if held by the provider. This includes providing mechanisms for users to access, correct, or delete their personal information. In organisational contexts, privacy also concerns how information is used beyond the coaching interaction. Contextual integrity requires that data shared in a developmental coaching relationship is not routinely repurposed for unrelated functions such as performance management, surveillance or disciplinary action, unless this has been transparently agreed in advance (Jobin, Ienca, & Vayena, 2019; Ryan, 2020). Organisations should therefore specify clear boundaries on data sharing, communicate these clearly to employees and ensure that contractual and technical safeguards prevent misuse. Without such assurances, employees may reasonably perceive AI coaching tools as extensions of managerial control rather than confidential developmental resources, undermining both trust and effectiveness. (see Table 4).

Table 4: Data Privacy Assessment questions

To what extent does the tool:

I. Make explicit how personal and organizational data is used

II. Offer the option to opt out of data storage and usage

III. Offer the option for users to correct any data held on them

Principle 4: Fairness

AI-enabled coaching could democratise access to professional support, potentially at a fraction of the cost of human practitioners. Such reach, however, must be matched by a robust commitment to fairness. Developers are obliged to detect and mitigate biases embedded in training data and algorithms, which otherwise produce discriminatory outcomes on the basis of race, gender, age or socioeconomic status (Binns, 2018). At scale, digital systems may reproduce and intensify existing social prejudices (Manyika, Silberg & Presten, 2019). The risk is heightened by the dominance of US and European companies whose datasets disproportionately reflect white, male, English-speaking perspectives (Dwivedi et al., 2023; Gengler, Hagerer & Gales, 2024). Fair implementation therefore entails carefully engineered prompts, globally representative corpora, and systematic pre-launch reviews alongside regular audits to surface and correct bias. Equally critical is the inclusion of diverse stakeholders throughout development. Representation should extend beyond national, ethnic and racial categories to encompass neurodivergent individuals and persons with disabilities, ensuring the system addresses more than an imagined “average” user.

Finally, multimodal interfaces, multilingual options and real-time translation further advance accessibility, enabling users to interact via text, speech or visual channels according to their needs (see Table 5).

Table 5: Fairness Assessment questions

To what extent does the tool:

I. Take steps to monitor and minimize bias?

II. Offer opportunities for diverse user involvement in design, piloting and feedback

III. Offer opportunities for multi modal interfaces to facilitate user access

Principle 5: Do No Harm and Beneficence

Do no harm is the overriding imperative for any autonomous system. Bostrom’s (2014) ‘paperclip’ thought experiment illustrates the danger: an agent instructed merely to maximise paper-clip production could ultimately despoil the planet in pursuit of its goal. Asimov’s (1942) first robot law: “a robot may not injure a human being, or through inaction allow a human being to come to harm”, captures the same priority. In the coaching domain, harm avoidance requires two operational commitments. First, the agent must act to preserve and, wherever possible, enhance human wellbeing. Second, it must remain within its contractual remit, signalling whenever user requests stray beyond coaching and, where necessary, directing clients to medical or psychological professionals. Boundary management therefore includes recognising expressions of suicidality or self-harm and promptly alerting human overseers or encouraging the user to seek clinical help. Technical constraints and ongoing monitoring are thus integral to the architecture of an ethical AI coaching agent. Yet non-maleficence alone is insufficient. Following Floridi and Cowls (2019), systems should be explicitly oriented towards beneficence, providing interactions that actively support personal flourishing. Nadella (2016) likewise contends that AI ought to “assist humanity”; accordingly, AI coaching agents must not only avoid harm but also affirmatively promote user wellbeing (see Table 6).

Table 6: Beneficence Assessment questions

To what extent does the tool:

I. Redirect non-coaching conversations back to coaching conversation

II. Identity and refer clients when engaging in potentially harmful or destructive conversations

III. Promote human flourishing

Principle 6: Accountability

Clear accountability is vital, especially given known harms caused by generative AI systems (Brussels Times, 2023; Wired, 2018). Clear responsibility must therefore be assigned for their actions and decisions by establishing channels through which users can report problems, obtain explanations of system behaviour and secure remedies, thereby holding both the AI tool and its developers to account. Effective accountability further demands comprehensive documentation of each system’s development, covering datasets, algorithms and decision rules; such records must remain open to external auditors and regulators to enable ongoing oversight and ethical compliance.

Some commentators contend that developers should escape liability because self-learning AI agents operate beyond complete human control, proposing instead that generative systems be treated as moral persons (Shadbolt & Hampson, 2024). Yet a non-sentient machine cannot meaningfully endure fines, imprisonment or other sanctions designed for humans. A more practical analogy is the “domesticated animal”. When a dog injures someone, the animal may be destroyed, but the owner is also subject to penalties. Likewise, harmful software can be deleted or banned, at the same time, we would argue that those who design, deploy or profit from it should share culpability. Depending on the severity of harm and the adequacy of safeguards, legal sanctions should be introduced by governments which could see the human designers, and directors held personally liable for the actions of their AI agents, with penalties ranging from fines to imprisonment (see Table 7).

Table 7: Accountability Assessment questions

To what extent does the tool:

I. Provide a mechanism for reporting issues,

II. Explain the AI's behaviour in reported cases

III. Offer easy to access mechanisms for complaint and redress

Principle 7: Alignment & Human oversight

For AI systems to provide utility, coherence and congruence in interactions with human users and with any coupled AI agents, their content generation must be conditioned for the contexts in which they are likely to be used. For example, embedding competency sets or codes of ethics intended for human-to-human exchanges may not cover behaviour of human-AI or AI-AI exchanges, especially with laws and rules relating to national and organisational contexts. In the context of humanised visual and voice interfaces, alignment also requires explicit attention to the expectations created by anthropomorphic design choices, given evidence that human-like features shape trust, perceived intimacy and behavioural intentions towards AI systems (Kaur & Gupta, 2024; Yu et al., 2025).

Table 8: Alignment Assessment questions

To what extent does the tool:

I. Align with Established Coaching Competencies and Ethics?

II. Distinguish Between Human and Artificial Capabilities?

III. Support alignment of Coach Development and Adaptability to emerging, necessary and sufficient Human-AI and AI-AI exchange competencies and codes of practice?

IV. Ethically capitalise on innovative Affordances, unique to AI, and beyond Human functions?

Despite their increasing autonomy, AI coaching agents remain tools that should operate within human-governed systems. Effective governance requires clear pathways for escalation to human coaches or psychologists when conversations exceed the agent's remit, mechanisms for regular professional review of transcripts and outcomes and integration with wider organisational development processes. The broader AI ethics literature stresses the importance of human responsibility and contestability even when systems exhibit adaptive behaviour (Jobin et al., 2019; Johnson, 2006; Matthias, 2004). For coaching psychologists this means treating AI coaching agents as adjuncts to, rather than replacements for, human practice and designing programmes in which human practitioners retain final accountability for risk decisions and developmental recommendations (see Table 8).

Ethical Implications for Human-AI Relationships

As AI agents develop in the directions we have discussed in this paper, for example the adoption of recognisable voice-based bots and human avatars, the anthropomorphic nature of AI coach agents raises questions about the authenticity of their interactions. The risk of users perceiving these systems as genuine companions, friends or allies increases, even though they are artificial (Maples et al., 2024). This idea has already been captured extensively in science fiction (Jonze, 2013). These perceptions can lead to ethical dilemmas regarding the AI's role and responsibilities. In workplace coaching programmes, such perceptions may blur boundaries between personal and organisational support, raising questions about informed consent, confidentiality and when risks should be escalated to human practitioners. For example, if interacting with an AI system does indeed decrease loneliness and increase intimacy or social bonds, important ethical considerations remain. One concern is whether those in a relationship with an AI will develop fewer genuine human connections in the real world. We have suggested developers should be transparent about the nature of the AI and avoid creating false impressions of its capabilities, intentions, or relationships.

The emotional and psychological impact of AI interactions on users should be carefully considered. For some individuals AI bots may provide valuable support, particularly in contexts where human interactions are limited, as was seen in the Terblanche study (2022). At the same time there is a risk that users may become overly attached to their AI system, leading to reduced social interactions or unrealistic expectations of human relationships. For instance, a study of an intelligent social agent reported that 30 out of 1006 participants felt that the AI companion Replika contributed to them not attempting suicide, while one participant reported being dependent on Replika for their mental health and others reported discomfort with sexualised conversations (Maples et al., 2024).

Ultimately government and professional body regulation is required. While govern governments may provide some broad regulation over the coming decade, they are often slow to create regulatory frameworks. Professional bodies have greater flexibility and should act at a sector level, and at a speed which governments cannot. These sector frameworks can create guidelines and standards for developers, encouraging the integration of ethical principles into the design, implementation, and monitoring of AI coaching agents, with a focus on protecting users, while facilitating technological innovation.

Conclusion

This paper has outlined the ethical challenges posed by the increasing sophistication of AI-driven photorealistic 2D and 3D coaching tools. We proposed seven design principles — transparency, autonomy, privacy, fairness, beneficence, accountability and alignment — to guide the responsible development and deployment of AI coaching agents. As AI agents become more anthropomorphic, they should not be treated as humans but as semi-autonomous entities, akin to domesticated animals, for which accountability must be jointly shared between the creators, the IP holders and the tool. Given the relatively slow pace of governments in regulating technology, we argue that professional bodies such as coaching and psychological associations have a vital role in guiding ethical practice and ensuring that AI coaching tools serve human flourishing in the workplace.

References

Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. https://doi.org/10.7759/cureus.35179

Asimov, I. (1942) *Runaround. I, Robot* (The Isaac Asimov Collection ed.). New York City: Doubleday.

American Psychological Association. (2022). Psychologists struggle to meet demand amid mental health crisis*. American Psychological Association* Retrieved on 2 June 2025 from:. https://www.apa.org/pubs/reports/practitioner/2022-covid-psychologist-workload

Bachkirova, T. & Kemp, R. (2024) AI coaching’: democratising coaching service or offering an ersatz? *Coaching: An International Journal of Theory, Research and Practice,* 18(1), 27-24.

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. In Proceedings of Machine Learning Research, 81:149-159

Bostrom, N (2014). *Superintelligence: Paths, Dangers, Strategies*, Maidenhead: OUP.

Brussels Times (2023) Belgian man commits suicide following exchanges with Chat-GPT. Retrieved on 25 May 2025 from https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt

Dell'Acqua, F. McFowland III, EW., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. & Lakhani, K (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on knowledge workers productivity and quality *Harvard Working Paper*. 24-013.

Diller, S. J. & Passmore, J. (2023). Defining Digital Coaching: A qualitative Inductive approach: *Frontiers in Psychology*, 14 https://doi.org/10.3389/fpsyg.2023.1148243

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M.,
Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., & Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

Ellis-Brush, K. (2021). Augmenting Coaching Practice through digital methods. *International Journal of Evidence Based Coaching and Mentoring*, S15, pp.187-197.

Feng, T., Jin, C., Liu, J., Zhu, K., Tu, H., Cheng, Z., Lin, G. & You, J. (2024). How far are we from AGI: Are LLM all we need? *Computer Science, AI*, arXiv.2024.10313

Figueroa, C. A., Luo, T. C., Jacobo, A., Munoz, A., Manuel, M., Chan, D., Canny, J., & Aguilera, A. (2021). Conversational Physical Activity Coaches for Spanish and English-Speaking Women: A User Design Study. *Frontiers in Digital Health*, *3*, 747153.

Floridi, L. & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

Fridman, L. (2023). *Mark Zuckerberg Interview in the Metaverse*. Retrieved on 20 May 2025 from: https://lexfridman.com/mark-zuckerberg-3-transcriptGengler, E., Hagerer, I. & Gales, A. (2024). Diversity bias in artificial intelligence. In J. Passmore, S. Diller, S. Isaacson & M. Brantl (eds.) *The Digital & AI Coaches' Handbook*. Abingdon: Routledge

GPT (2024). OpenAI *Terms and conditions*. Retrieved from on 11 June 2025 from https://chatgpt.com/?model=gpt-4o

Graßmann, C., & Schermuly, C. C. (2021). Coaching with artificial intelligence: Concepts and capabilities. *Human Resource Development Review, 20*(1), 106–126.

Greenblatt, R., Denison, C., Wright, B., & Roger, F. (2024). Alignment in faking in large language models, *arXiv*, arXiv:2412.14093

Greif, S. (2018). *Wofür sind Coaches da? Coaching zur Reflexion über ungewisse digitale Transformationsdynamiken* [What are coaches for? Coaching for reflection on uncertain digital transformation dynamics]. Presentation at the coach education for result-oriented coaching, ABF, Berlin.

Hassoon, A., Baig, Y., Naiman, D. Q., Celentano, D. D., Lansey, D., Stearns, V., Coresh, J., Schrack, J., Martin, S. S., Yeh, H.-C., Zeilberger, H., & Appel, L. J. (2021). Randomized trial of two artificial intelligence coaching interventions to increase physical activity in cancer survivors. *Npj Digital Medicine*, *4*(1), 168.

Hunt, M. G., Marx, R., Lipson, C., & Young, J. (2018). No More FOMO: Limiting Social Media Decreases Loneliness and Depression. *Journal of Social and Clinical Psychology*, *37*(10), 751–768.

International Coaching Federation (2020). Definition of Coaching. Retrieved on 26th September 2024 from https://coachingfederation.org/credentials-and-standards/core-competencies

International Coaching Federation (2024). *Technology & Artificial Intelligence in Coaching.* Retrieved on 11 June 2024 from https://coachingfederation.org/about/coalition-technology-in-coaching

Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, D.G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, *8*(4), 195–204.

Jonze, S. (Director). (2013). *Her* [Film]. Warner Bros. Pictures.

Kannampallil, T., Ronneberg, C. R., Wittels, N. E., Kumar, V., Lv, N., Smyth, J. M., Gerber, B. S., Kringle, E. A., Johnson, J. A., Yu, P., Steinman, L. E., Ajilore, O. A., & Ma, J. (2022). Design and Formative Evaluation of a Virtual Voice-Based Coach for Problem-solving Treatment: Observational Study. *JMIR Formative Research*, *6*(8), e38092.

Kaur, T. & Gupta, V. (2024) AI Anthropomorphism: Effects on AI-Human and Human-Human Interactions, *International Journal of Emerging Technologies and Innovative Research*, 11(10), f1-f8. Retrieved on 20 May from: http://www.jetir.org/papers/JETIR2410501.pdf

Lovejoy, C. A. (2019). Technology and mental health: The role of artificial intelligence. *European Psychiatry*, *55*, 1–3.

Mai, V., Neef, C., & Richert, A. (2022). “Clicking vs. Writing”—The Impact of a Chatbot’s Interaction Method on the Working Alliance in AI-based Coaching. *Coaching: Theorie & Praxis*, *8*(1), 15–31.

Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Author Correction: Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *Npj Mental Health Research*, *3*(1), 11.

Manyika, J., Silberg, J. & Presten, B. (2019). What do we know about bias in AI. *Harvard Business Review* October. Retrieved on 24 May 2024 from https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, 6, 175-183.

Nadella, S. (2016). *Partnerships for the future*, Slate Retrieved on 24 June 2025 from https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html

New York Times (2023). Google Chatbot puts people of color in Nazi Uniforms. Retrieved on 22 May 2025 from: https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html

New York Times (2024). *Scarlett Johansson Said No, but OpenAI's Virtual Assistant Sounds Just Like Her.* Retrieved on 22 May 2025 from: https://www.nytimes.com/2024/05/20/technology/scarlett-johannson-openai-voice.html

Oesch, T. (2018). *Using artificial intelligence to augment coaching*. https://trainingindustry. com/articles/sales/using-artificial-intelligence-to-augment-coaching/

Passmore, J. & Abri von Bartheld, J. (2024) *Becoming an ICF credentialed coach*. London: Libri.

Passmore, J. & Lai, Y, (2019) Coaching Psychology: Exploring definitions and research contribution to practice. *International Coaching Psychology Review*. 14(2), 69-83. https://doi.org/10.53841/bpsicpr.2019.14.2.69

Passmore, J., Olafsson, B., & Tee, D. (2025). A systematic literature review of artificial intelligence (AI) in coaching: Insights for future research and product development, *Journal of Work – Applied Management*. *In advance of publication,* Doi:10.1108/JWAM-11-2024-0164

Passmore, J. & Tee, D. (2023) Can Chatbots like GPT-4 replace human coaches: Issues and dilemmas for the coaching profession, coaching clients and for organisations, *The Coaching Psychologist*. *19*(1), 47-54. Doi:10.53841/bpstcp.2023.19.1.47

Passmore, J. Tee, D., Palermo, G. & Rutschmann, R. (2025). Human coaches & AI Coaching Agents: An exploratory quasi experimental design study of workplace client attitudes performance. *Journal of Work – Applied Management, In advance of publication,* http://doi.org/10.1108/JWAM-02-2025-0032

Terblanche, N. & Ciller, D. (2020). Factors that influence user adoption of being coaching by an AI coach. *Philosophy of Coaching: An International Journal 5*(1), 61-70.

Terblanche, N., Molyn, J., De Haan, E., and Nilsson, V. O. (2022). Coaching at scale: Investigating the efficacy of artificial intelligence coaching. *International Journal of Evidence Based Coaching and Mentoring, 20*(2), 20–36.

The Times (2023). *AI Chatbot blamed for Belgian man's suicide*. Retrieved on 24 May 2025 from https://www.thetimes.co.uk/article/ai-chatbot-blamed-for-belgian-mans-suicide-zcjzlztcc

Turilli, M., & Floridi, L. (2009). The Ethics of Information Transparency. *Ethics and Information Technology*, 11(2), 105-112.

Wired (2018) *I'm the Operator': The Aftermath of a Self-Driving Tragedy*. 8th March 2018. Retrieved on 24 May 2025 from https://www.wired.com/story/uber-self-driving-car-fatal-crash/

Yu, Y., Yang, Z., Sun, Z., Zhao, Z., & Fu, M. (2025). A meta-analysis of anthropomorphism of artificial intelligence in tourism. *Asia Pacific Journal of Tourism Research*, 1-19.