Topoi
https://doi.org/10.1007/s11245-025-10245-w

AI Companions for Philosophical Health: a Human-in-the-Loop Framework

Luis de Miranda[1,2,3,4]

Accepted: 24 July 2025

Abstract
This article explores the reciprocal relationship between artificial intelligence and philosophical health – defined as eudynamic adequation between thoughts and actions. Rather than simply examining how AI might enhance philosophical practice, we investigate how philosophical health can enhance AI applications while preserving the human character of philosophical reflection. We introduce the C.I.P.H.E.R. model (Crealectic Intelligence and Philosophical Health for Eudynamic Realities), a novel framework allowing for AI and human philosophical capacity to collaborate through a "human-in-the-loop" approach. Drawing on original data from a survey of 288 participants conducted between July 2023 and March 2025, we examine contemporary philosophical perspectives across six dimensions: bodily sense, sense of self, belonging, possibility, purpose, and philosophical worldview. Key findings reveal significant patterns including widespread loneliness despite connection (30.6%), evolving self-perception (75.3%), and gaps between purpose recognition and implementation. These empirical insights are evaluated against our theoretical framework for mutual enhancement between philosophical health and AI applications. We address fundamental challenges of technological mediation in philosophical inquiry, particularly regarding embodied understanding and authentic meaning-making.

Keywords Philosophical health · Artificial intelligence · Crealectic intelligence · AI ethics · Cogeneration · Human in the loop · Philosophical counselling · Digital philosophy

1 Introduction

Know Thyself: this ancient imperative assumes unprecedented complexity as neural networks simulate aspects of human cognition with increasing sophistication. Philosophical health – defined as personal coherence between world-conception, thoughts, values, and actions – traditionally requires embodied dialogue and lived experience (de Miranda 2023). Yet increasing numbers of individuals engage AI-based conversational agents for existential inquiry, necessitating critical examination: can artificial intelligence enhance our pursuit of philosophical health and authentic self-understanding?

Skeptics including Dreyfus (1992) and Turkle (2011) warn that technological mediation risks compromising the intuitive connections essential to philosophical insight. Conversely, empirical research suggests AI analysis of digital behavior sometimes reveals personality traits with greater accuracy than human judgment (Youyou et al. 2015), indicating machines may detect patterns in human expression that escape conscious awareness. Floridi (2014) argues computational systems provide novel conceptual frameworks challenging entrenched thinking patterns, potentially expanding philosophical reflection beyond traditional limitations.

This article explores the thesis that AI-assisted philosophical practice may – under certain conditions – enhance our capacity for meaning-making and authentic self-understanding. To ground our argument, we present original empirical data from a Philosophical Health International

✉ Luis de Miranda
luis.demiranda@utu.fi; luis.demiranda@uu.se

1 Department of Philosophy, University of Turku, Turku, Finland

2 Turku Institute for Advanced Studies (TIAS), 20014 Turku, Finland

3 Centre for Research Ethics and Bioethics, Uppsala University, 75310 Uppsala, Sweden

4 Center for Wellbeing, Welfare and Happiness, Stockholm School of Economics, 11383 Stockholm, Sweden

Published online: 21 August 2025

survey examining contemporary philosophical perspectives across six dimensions of the SMILE_PH method (Sense Making Interviews Looking at Elements of Philosophical Health). Our findings from 288 participants provide quantitative support for theoretical approaches and foundational insights for AI applications in philosophical health.

Finally, we introduce the C.I.P.H.E.R. model (Crealectic Intelligence and Philosophical Health for Eudynamic Realities), a novel framework that helps to reimagine human-artificial intelligence relationships. Positioning AI as collaborative partner rather than autonomous guide, our critical "anthrobotic" approach (de Miranda et al. 2016) preserves philosophical inquiry's irreducibly personal dimension while exploring computational assistance in deep orientation and sense-making.

2 Philosophical Health as Sense-Making Paradigm

Philosophical health represents a distinct paradigm for human flourishing, complementing yet differing fundamentally from psychological well-being. While psychological approaches emphasize symptom reduction, pathological diagnosis, and functional adaptation, philosophical health addresses fundamental questions of meaning, coherence, and existential orientation.

We define philosophical health as eudynamic adequation between conceptualized thoughts and actions, which involves a spirit of compossibility with reality (de Miranda 2024a). Compossibility, derived from Leibniz's theory of possible worlds (Brown and Chiek 2016), refers to active harmony or non-contradiction between elements constituting coherent existence. This principle prevents what we term "philosophical illness" – the contradictory coexistence of incompatible worldviews, beliefs and practices within the same person or group.

This conceptualization aligns with ancient traditions viewing philosophy as life practice oriented toward human realization (Hadot 1995), notably Aristotle's eudaimonia emphasizing potential actualization through virtuous activity and practical wisdom. While psychological health focuses on improved functioning within existing frameworks, philosophical health interrogates the frameworks themselves: What constitutes flourishing? How should one relate to others and world? What personal cosmologies should guide action? These questions transcend symptom management, engaging fundamental meaning structures orienting human existence.

The SMILE_PH method provides a structured exploration of philosophical health through six interconnected dimensions:

Bodily Sense How do I relate to my physical existence and embodied experience? This dimension examines one's relationship with embodied existence, recognizing Merleau-Ponty's (2012) "lived body" as experiential foundation. Research in embodied cognition confirms that physical embodiment fundamentally shapes understanding and values (Johnson 2007).

Sense of Self Who do I think I am and how do I understand my identity? This dimension explores one's personal concept of the self, including questions of authenticity and self-knowledge. Contemporary theories emphasize identity's dynamic, narrative nature rather than static conception (Gallagher 2000; Zahavi 2005).

Sense of Belonging Where and with whom – or what – do I feel connected? This dimension investigates how individuals establish connections with others, communities, and broader ensembles or realities. This addresses Heidegger's (1962) "being-with" (Mitsein) and social ontology (Searle 2010). Research confirms "well-belonging" as fundamental to human flourishing (James et al. 2014; de Miranda 2020a).

Sense of the Possible What futures can I desire, imagine and create? This element examines how individuals relate to potential futures and imagine alternative realities. Drawing on existential philosophy's emphasis on possibility and project (Sartre 2007), this dimension explores how possibility perception shapes current experience, feeling and action.

Sense of Purpose What gives my life meaning and long-term direction? This dimension explores how individuals establish meaning and symbolic orientation in their lives. This engages Frankfurt's (1988) "volitional necessity" and Korsgaard's (1996) "practical identity" – descriptions under which we find life worth living and actions worth undertaking for the long-term.

Philosophical Sense What is my overall framework or personal cosmology for understanding reality? This meta-dimension examines worldview and conceptual frameworks, allowing integration of other elements while maintaining awareness of theoretical foundations.

SMILE_PH applications in various contexts demonstrate promise in clinical and non-clinical settings (de Miranda, Levi & Divanoglou 2023). The philosophical orientation of this method distinguishes it from purely psychological approaches through focus on conceptual meaning over biographical narration, universalizable values over reactive coping, and open discourse over symptom detection.

The quantitative investigation presented in the following section is an exploration into experimental philosophy. It provides insights into contemporary philosophical perspectives across the six dimensions of philosophical health. These preliminary findings are an example of collective data that may inform our investigation about AI integration in philosophical practice.

3 Quantitative Investigation: the PHI Survey

This section presents methodology and findings from a Philosophical Health International (PHI) survey conducted between July 2023 and March 2025, providing empirical context for AI-enhanced philosophical health exploration.

3.1 Methodology

The survey employed cross-sectional, quantitative design exploring multiple possibilities within each philosophical health element, collecting responses from 288 anonymous participants across bodily sense, sense of self, belonging, possibility, purpose, and philosophical worldview dimensions. Each category presented 10 statements; participants could select all applicable responses. An "Other" option bridged quantitative and qualitative approaches, allowing contributions beyond pre-defined options.

Distribution occurred through the Philosophical Health International website and social media channels, employing convenience sampling by attracting individuals with philosophical interest. Data collection continued over 20 months, with automatic anonymous recording through the platform, and percentage calculations performed for each statement category. No personal identifying information was collected, encouraging candid responses to potentially sensitive philosophical questions.

Participants received this introduction: "The following questions were designed to help you reflect on your philosophical health, based on the SMILE_PH method developed by Dr Luis de Miranda. This anonymous introduction helps you decide if you wish to become philosophically healthy. By answering, you agree that overall percentage results may be used for research purposes. You will not leave your email, and we cannot see which person answered what. Options are limited and constrained; helpful philosophical health work requires face-to-face dialogue through which you elaborate your own answers. Remember: there is no unique way of reaching philosophical health, but many singular and diverse ones."

This dataset represents original research synthesized exclusively for this study. The 288 responses synthesized for this investigation have not been previously published, analyzed, or used in other research publications. The results show strong experimental evidence that the six elements of philosophical health (body, self, belonging, possibility, purpose and philosophical sense) are valid and necessary constructs.

3.2 Key Findings

Bodily Sense Only 56.2% agreed "My body is mostly my friend," suggesting widespread bodily alienation aligning with contemporary embodiment research (Tylka and Wood-Barcalow 2015). This finding highlights complex mind-body relationships requiring integrated embodiment dialogue in philosophical health interventions.

Sense of Self 75.3% believed their self is evolving, aligning with lifelong identity development theories (Kroger and Marcia 2011). However, only 12.5% considered themselves "the highest version of my possible selves," indicating widespread unrealized potential perception or epistemic humility. Discussions about the self are therefore needed in philosophical health dialogue.

Sense of Belonging 30.6% reported loneliness despite connection forms, aligning with global social isolation concerns (Cacioppo and Cacioppo 2018). Additionally, 33% agreed "It is better to belong to a purpose rather than a group," suggesting shifts toward purpose-driven social connection. Personal well-belonging clearly needs to be discussed in philosophical health interventions.

Sense of the Possible While 62.2% viewed futures as possibility-filled, only 28.8% believed they decide what is possible, suggesting perceived agency deficits with motivation implications (Bandura 1977). 44.8% agreed "Everything is possible," indicating high openness to ultimate possibility, but not necessarily a possible that can be controlled.

Sense of Purpose 47.6% reported high purpose sense, yet only 38.9% agreed their life mission shapes existence, indicating gaps between purpose recognition and implementation, reflecting existential concerns about authenticity and commitment (Sartre 2007). These results emphasize the importance of a well-defined sense of purpose in philosophical health interventions.

Philosophical Sense 70.5% agreed "Understanding improves our actions," indicating belief in interpretation's practical value. 33.3% viewed the universe as "a sort of mind," suggesting openness to panpsychist or idealist perspectives (Goff 2019) and supporting theory-practice adequation possibilities. The pertinence of the core definition

of philosophical health as alignment between thought and practice is confirmed by the results.

4 Digital Knowledge Infrastructures in Philosophy

How relevant are digital approaches in philosophical practice?

Philosophy traditionally centers on face-to-face dialogue, physical texts, and embodied pedagogical practices. Digital humanities projects transform philosophical text access, analysis, and interpretation (Berry and Fagerjord 2017), though philosophy remains underrepresented compared to other disciplines. Ess (2004) notes philosophical engagement with digital technologies emphasizes critical analysis of social and ethical implications rather than embracing tools for philosophical practice, reflecting what Brey (2000) terms "technological ambivalence."

Computational platforms offer opportunities. Digital accessibility addresses geographical, financial, and social barriers to philosophical counselling (Keegan 2012). Consider, for instance, rural factory workers who experience existential anxiety: accessing human counselling requires extensive travel and prohibitive costs. Digital assistance may democratize philosophical reflection access, aligning with traditions viewing philosophical inquiry as universal right rather than elite privilege. In this spirit Philosophical Health International proposes since June 2025 a free access to an AI-chatbot, Philai, trained in the SMILE_PH method (https://www.philosophical.health/).

Digital approaches also enable unprecedented collective data collection and analysis, potentially revealing invisible patterns across populations. These quantitative approaches complement philosophical inquiry's traditionally qualitative nature, offering "epistemological pluralism" (Brey 2000) – multiple ways of knowing. AI systems identify patterns escaping human perception, revealing relationships between philosophical well-being elements informing integrated approaches.

Longitudinal tracking of philosophical development becomes possible through digital platforms. Peters and Jandrić (2018) argue that traditional approaches to philosophical education have underexplored the dynamic, evolving nature of philosophical development in favor of examining fixed philosophical positions or states. Digital platforms may capture this evolution, providing insights into philosophical health dynamics rather than static dimensions.

However, digitization presents significant challenges in studying eudaimonic well-being. Dreyfus (2001) argues aspects of human understanding resist formalization and digitalization. Tacit, embodied, contextual dimensions of philosophical knowledge – Polanyi's (1966) "personal knowledge" – may be diminished in digital translation. Digital approaches risk reducing philosophical health to quantifiable metrics, sacrificing depth for scale (Borgmann 1984).

Digital platforms mediate philosophical experience in ways fundamentally altering its character. Ihde (1990) argues that technological mediation transforms rather than transmits experience. Digital environments may encourage fragmented attention and consumption over contemplation (Carr 2010), opposing philosophical inquiry's deliberate, reflective nature.

These challenges suggest that digital approaches to philosophical health must be developed with careful attention to their limitations and potential distortions. The goal should not be to simply digitize existing philosophical practices but to thoughtfully reimagine philosophical health in the digital context while preserving its essential qualities. While digital platforms offer unprecedented access and pattern-recognition capabilities, the challenges of embodiment, depth, and authenticity require careful design consideration. The following framework addresses these tensions by proposing a collaborative model preserving the irreducibly human aspects of philosophical development.

5 The C.I.P.H.E.R. Model

The C.I.P.H.E.R. model (Crealectic Intelligence and Philosophical Health for Eudynamic Realities) is a novel framework that allows us to evaluate the integration of artificial intelligence into philosophically healthy practice. This section details the model's theoretical foundations, architecture, and practical implementation, drawing on both established philosophical traditions and findings from the PHI survey.

The CIPHER model (see Fig. 1) emerges from the intersection of two distinct but complementary theoretical frameworks: philosophical health and crealectic intelligence. Crealectic intelligence represents a form of thinking that transcends both analytical and dialectical modes to engage with creative potential and the composition of emergent possibilities (de Miranda 2020b).

The term "crealectic" combines "creative" with "logos," suggesting a form of thinking that emphasizes multiplicity and generative potential rather than mere analysis or simple binary oppositions (de Miranda 2020b). Unlike analytic intelligence, which excels at decomposition and systematic examination, or dialectic intelligence, which focuses on the resolution of contradictions, crealectic intelligence emphasizes the creative generation of new compossibilities – compatible possibles – and inner innovation. This approach aligns for instance with how some architects working on urban planning problems don't simply analyze

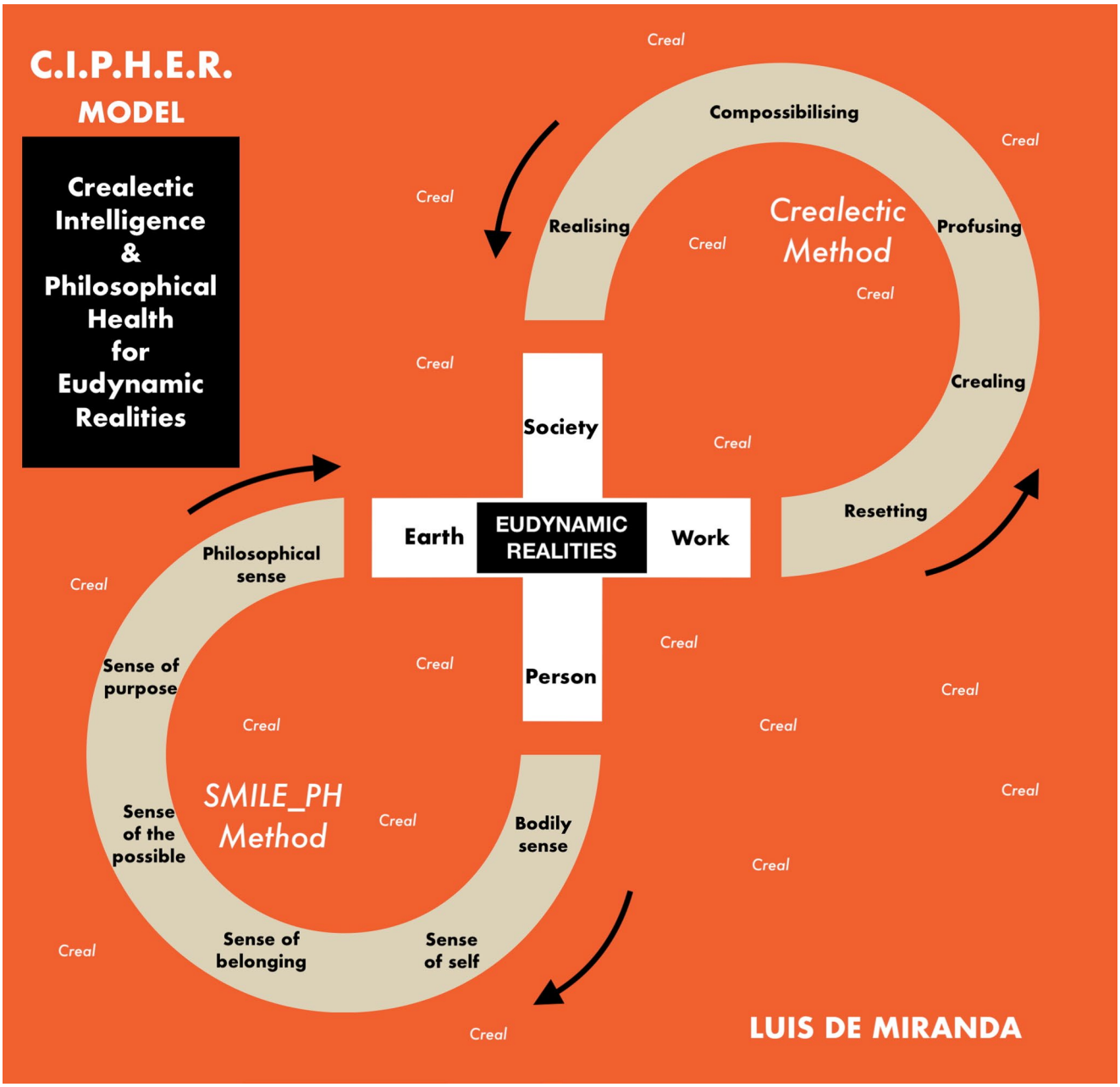

Fig. 1 The C.I.P.H.E.R. model. (source: Luis de Miranda). Original conceptual framework developed by the author

existing traffic patterns (analytical) or resolve the contradiction between cars and pedestrians (dialectical), but instead creatively generate new spatial possibilities that allow multiple and even unpredictable transportation modes, green spaces, and community gathering areas to coexist and evolve eudynamically, in allowing for a good relationship with the possible.

The philosophical foundation of crealectic intelligence lies in the "Creal", a conceptual abbreviation for Creative Real (de Miranda 2017), representing the dynamic generative potential inherent in reality. Crealectic intelligence distinguishes itself through its cogenerative orientation, focusing on awakening inner and outer compossibilities rather than simply analyzing existing patterns. For instance, a neighborhood struggles with both housing costs and social isolation. Instead of treating these as separate problems requiring distinct solutions, compossibilising recognizes them as expressions of a single possibility seeking emergence. Co-housing experiments emerge for instance where residents share resources and decision-making, simultaneously reducing individual costs while creating deeper community bonds. What seemed like two obstacles transforms

into a cogenerative tension that births a renewed form of urban living, opening the space for further compossibilities.

The CIPHER model combines the SMILE_PH method with the Crealectic Method (de Miranda 2024). The Crealectic Method is a five-step approach to foster a "compossible mindset", a way of thinking that ensures diverse possibilities can eudynamically coexist in the same world. The method combines creative, real, and meta-analytical thinking to address our interconnected planet's challenges where local innovations can sometimes lead to global problems. The five steps of the Crealectic Method are:

Resetting "I decide for a moment that I know nothing, except that I feel alive." Deliberately clearing the mind of preconceptions, biases, judgments and cognitive clutter, in order to create, as much as possible, a space of personal availability. This mental reset, once repeated as spiritual exercise, helps break free from habitual patterns and opens one to the possibility of new possibilities. Though perfect resetting is difficult, even attempting it provides benefits.

Crealing "The universe creates through me." A blend of "creation" and "feeling" that connects individuals with the ongoing creative processes within and around them. This step involves actively participating in the "creative élan vital" or life force, recognizing we are not just observers but "intercreators" constantly shaping and being shaped by the universe of our lived worlds.

Profusing "I let possibilities emerge without judgment." Allowing ideas to flow freely without censorship or imposed hierarchies. This step celebrates unbridled creativity where thoughts merge, mingle, and bloom with uninhibited vigor. It taps into the brain's default mode network, strongly associated with dreaming, enabling richer and more diverse ideation.

Compossibilizing "I weave multiple possibilities into harmony." Harmonizing diverse possibles to coexist in dynamic synergy. This practice weaves multiple possibilities into a coherent, integrated network, finding unity in diversity. It involves reconciling seemingly conflicting or disparate ideas while also accepting that not everything can coexist.

Realizing "I understand and make it real." This dual-natured step involves both understanding something clearly and bringing ideas to fruition. It connects cognitive insights with tangible action, translating abstract concepts into pragmatic outcomes aligned with deeper purposes.

The method has been applied in organizations like Vattenfall's R&D unit to foster innovation that balances economic viability with environmental responsibility. The Crealectic Method ultimately aims to help approximate "the best of all compossible worlds" by ensuring that our innovations and actions contribute to sustainable, harmonious systems.

Future development of the CIPHER model must be informed by continuous dialogue between philosophical theory, empirical research, and engagement with practical implementation, ensuring that the model evolves within a deeper understanding of philosophical health and existential creativity. The aim is Enriched – or Eudynamic – Realities, at the intersection of person, earth, work and society.

By combining theoretical care with practical implementation, the CIPHER model offers a framework for enhancing philosophical reflection while acknowledging the irreducibly creative dimensions of reality. Any technological augmentation of philosophical practice must preserve and enhance human agency, decision-making and its domain of compossibility. This earthly conscious human-in-the-loop approach aims to achieve this balance, positioning computational thinking as a partner rather than a replacement in the pursuit of eudynamic, enriched realities.

The system provides what Pea (2004) calls "scaffolding" for meta-cognitive processes, supporting users in reflecting on their own thinking without prescribing specific conclusions. This approach enhances what Schön (1983) terms "reflection-in-action" – the capacity to think about one's thinking during the process of engagement. The goal resembles what Engelbart (1962) called "intelligence augmentation" rather than artificial intelligence, using technologies of the self to enhance distinctively human capabilities rather than simulating or replacing them. A human-in-the-loop (HITL) methodology also builds on Lara and Deckers' (2019) critique of "exhaustive enhancement" systems that would make decisions for humans, instead examining the conditions of possibility for anthrobotic Socratic assistants, in which answering the questions is doubled by questioning the answers.

The effective implementation of a human-in-the-loop approach (HITL) requires clear allocation of roles between digital systems and human practitioners. This allocation should reflect the distinctive strengths and limitations of both human and artificial intelligence, creating what de Miranda, Ramamoorthy, and Rovatsos (2016) call a healthy "anthrobotic" partnership, a symbiotic or cogenerative collaboration between human and technological capabilities.

The AI component can systematically generate alternative viewpoints and possibilities that extend beyond the human practitioner's current thinking. Boden (2004) argues that computational systems produce what she calls "transformational creativity" by exploring conceptual spaces in parallel ways. For example, an AI system might systematically explore the concept of "justice" by examining it

through unexpected lenses, combining it with concepts from biology (symbiotic justice), physics (conservation of moral energy), or music theory (harmonic justice).

Human practitioners should retain primary responsibility for determining what matters and why. As Taylor (1989) emphasizes, strong evaluation requires engagement with constitutive goods that define one's identity and orientation. The CIPHER model recognizes that these deeply personal evaluations require human agency and cannot be delegated to technological systems. Each human carries both a singular and a universal perspective on reality enrichment. That is the *Person* concern at the center of the infinite CIPHER loop. This personal concern is embodied.

Human practitioners also excel at understanding the rich contextual nuances that frame their engagement with the world. As Dreyfus (1992) argues, human understanding is inherently situated in a context of practices, an embodied and enacted transformation of the environment. The CIPHER model acknowledges this contextual understanding as a human capacity that grounds philosophical creativity in practice. That is the *Work* presence at the center of the loop. This dimension is enactive: it arises through interaction.

Both human and AI components may participate in the ongoing philosophical dialogue, with the AI offering prompts, connections, and reflections while the human, in dialogue with other humans, guides the overall direction and evaluates the relevance and significance of the exchange. This shared responsibility reflects what Gadamer (1975) describes as the dialogical nature of understanding, where meaning emerges through the interplay of shared perspectives. That is the *Society* aspect of the infinite loop, by which we are all embedded in norms and collective discourses.

This role allocation recognizes what Russell (2019) identifies as a key principle for beneficial AI, the need to design systems that complement human and more-than-human capabilities rather than competing with them, within an extended space occupied by all living beings. By allocating responsibilities according to the distinctive strengths and vulnerabilities of living intelligenc and artificial intelligence, the CIPHER model creates not only what Licklider (1960) envisioned as a "man-computer symbiosis" but allows for the eudynamic emergence of a symbiotic planet. That is the *Earth* quality in the infinite CIPHER loop, representing the extended nature of cognition.

6 From Theory To Real-Life Applications: CIPHER Model in Practice

How may the CIPHER framework improve upon typical AI chatbot interactions for philosophical inquiry? This section demonstrates the model's practical advantages through three key areas identified in our aforementioned PHI survey findings.

6.1 Addressing the Purpose-Implementation Gap

Our survey revealed that 47.6% of respondents report having a high sense of purpose, yet only 38.9% agree that their mission shapes their daily existence. This gap suggests many people struggle to translate abstract purpose into lived practice.

A standard chatbot might ask "What is your purpose?" and provide generic advice about goal-setting or motivation techniques, treating this as a problem-solving exercise. The CIPHER model recognizes this as a philosophical rather than merely practical problem. The AI may draw for instance on frameworks like Frankfurt's (1988) analysis of deep desires and also evoke Korsgaard's (1996) work on practical identity to help users explore the relationship between their values and actions. Instead of offering solutions, it poses questions: "What makes this purpose feel disconnected from your daily choices?" The user maintains authority over interpretation while the AI provides philosophical frameworks for deeper reflection. The dialogue becomes philosophical – exploring meaning and coherence – rather than advisory.

6.2 Supporting Authentic Connection

Despite various forms of social connection, 30.6% of survey respondents reported feeling lonely, suggesting that mere social contact doesn't address deeper needs for belonging.

Standard responses might suggest joining groups, making more friends, or practicing social skills – treating loneliness as a behavioral problem. The CIPHER model recognizes loneliness as potentially indicating a need for authentic rather than merely social connection. A healthy AI may offer philosophical perspectives on different types of belonging – from Heidegger's (1962) "being-with" to Buber's (1970) I-Thou relationships – while the human evaluates which concepts resonate with their experience. The focus shifts from "how to be less lonely" to "what kind of connection do I actually need?" The conversation addresses existential rather than surface-level concerns about human connection.

6.3 Navigating Mind-Body Questions

Our survey revealed complex attitudes toward embodiment, with 43.1% agreeing "mind and body are one" and 19.1% explicitly endorsing separation, suggesting many people lack clear frameworks for understanding their embodied existence.

A chatbot might provide information about stress management, physical health, or mind-body techniques without addressing the underlying philosophical questions. The CIPHER model acknowledges the philosophical complexity of embodied existence. While recognizing AI's limitations in directly addressing bodily experience, it can present different philosophical approaches to embodiment – from phenomenology to embodied cognition research – while the human reflects on how these frameworks relate to their lived experience. The AI explicitly acknowledges what it cannot access: the felt sense of being embodied. The dialogue honors both the complexity of embodiment and the limits of AI understanding of what it is to be embodied as a person, extended on a earthly environment, enactive in work, and embedded in societies.

6.4 The Human-in-the-Loop (HITL) Difference

What distinguishes the CIPHER approach from standard AI interactions is the systematic preservation of philosophical and creative agency in all matters of ultimate significance. The AI contributes frameworks and pattern recognition while humans maintain authority over interpretation, evaluation, and implementation. This approach ensures that users engage in genuine philosophical reflection rather than consuming pre-packaged advice. As Gallagher (2020) notes, authentic philosophical development requires embodied dialogue and shared understanding that AI cannot replicate. The CIPHER model addresses this limitation by maintaining clear boundaries: AI assists with conceptual exploration while humans do the essential work of worldmaking meaning-making.

The model thus transforms AI from an advice-giver into what Stokes (2002) calls a "creative collaborator," expanding possibilities for reflection while preserving the irreducibly human dimensions of inquiry. The model thus represents a middle path between uncritical techno-optimism and categorical rejection of AI, seeking to enhance philosophical health and existential creativity while preserving an essentially life-affirming character.

The HITL approach maintains human agency as central while leveraging AI capabilities for pattern recognition, information integration, and perspective generation. The following section addresses the ethical framework necessary to guide this integration.

7 Ethical Framework for AI in Philosophical Health

The digitization of philosophical health raises significant ethical questions that extend beyond general concerns about technology ethics. These questions touch upon fundamental aspects of human autonomy, authenticity, authority, authorship, and the nature of philosophical understanding itself.

Privacy concerns take on special significance in philosophical health contexts. As Nissenbaum (2010) argues, privacy must be understood contextually, and philosophical inquiry represents a unique context with its own norms and expectations. The intimate nature of philosophical self-exploration demands special consideration of what Manders-Huits (2011) calls "moral identification", the ways in which personal data relates to and potentially affects personhood.

Questions of authenticity emerge when philosophical reflection is mediated by digital technologies. As Taylor (1991) argues, authenticity involves not just self-expression but engagement with "horizons of significance" that transcend individual preference. Digital philosophical practice must support connection to these broader horizons rather than encouraging what Borgmann (1984) calls "hyperreality," a simulated environment disconnected from intuitive significance or natural creativity.

Concerns about justice and access must also be considered. As van Dijk (2020) demonstrates, the "digital divide" encompasses not just access to technology but the skills, motivations, and opportunities to use it meaningfully. Ensuring that algorithmic resources do not exacerbate existing inequalities in authority and existential authorship requires attention to these multiple dimensions of digital justice.

Addressing these ethical considerations also requires what Floridi and Taddeo (2016) call "distributed responsibility," a recognition that ethical outcomes depend on multiple stakeholders and design choices throughout the development process. The CIPHER model aims to incorporate these ethical considerations – a care for the person, the earth, our ways of working and our societies – into its core design rather than treating them as external or secondary constraints.

Drawing on both philosophical ethics and empirical insights from the PHI survey, we propose four core principles to guide the responsible development and implementation of AI systems for philosophical health.

7.1 The Principle of Autological Autonomy (or Autonomous Self-Development)

The first principle concerns preserving and enhancing human autonomy in philosophical self-development.

"Autological" combines "auto" (self) with "logos" (reasoned discourse), emphasizing experiential and personal philosophical reflection. This principle directly addresses what Lara and Deckers (2019) identify as the central challenge of AI enhancement: ensuring that technological assistance enhances rather than replaces human moral agency. As O'Neill (2002) argues, genuine autonomy involves not merely freedom from constraint but the positive capacity for self-legislation through rational and existentia reflection. In the context of philosophical health, this means that AI systems should enhance rather than diminish users' capacity for independent philosophical judgment. The PHI survey findings reinforce the importance of autonomy in philosophical health. The substantial percentage of respondents (40.3%) who endorsed the statement "I belong to myself" suggests a widespread valuing of self-determination. Similarly, the finding that 75.3% of respondents view their self as evolving indicates a dynamic, developmental understanding of identity that requires ongoing autonomous engagement.

7.2 The Principle of Hermeneutic Integrity

The second principle concerns maintaining the depth and authenticity of philosophical interpretation. "Hermeneutic" refers to the art and science of interpretation, particularly the understanding of texts and experiences in their full context and significance. This principle holds that AI integration should preserve and enhance users' capacity for meaningful self-interpretation. As Ricoeur (1992) argues, hermeneutic understanding involves a "detour" through cultural symbols, narratives, and conceptual frameworks rather than direct introspective access. In philosophical health, this means that AI systems should support rather than short-circuit the interpretive process through which individuals make sense of their experiences and values. The PHI poll findings highlight the importance of interpretive processes in philosophical health. The strong endorsement (70.5%) of the statement "Understanding improves our actions" suggests widespread recognition of the practical significance of hermeneutic processes. Similarly, the diverse responses to philosophical worldview questions indicate the plurality of interpretive frameworks through which individuals make sense of their experiences.

7.3 The Principle of Philosophical Authenticity

The third principle concerns preserving and enhancing authentic engagement with philosophical questions. "Authenticity" here refers not merely to sincerity but to what Taylor (1991) calls "strong evaluation", an engagement with questions of what is truly worthy or significant beyond subjective preference. This principle holds that AI systems should support rather than standardize philosophical development. As Heidegger (1962) argues, authentic existence involves confronting fundamental questions about being and meaning rather than merely conforming to conventional patterns. In philosophical health, this means that AI systems should support genuine philosophical questioning rather than guiding users toward predetermined conclusions or standardized positions. The PHI poll findings suggest the importance of authenticity in philosophical health. The diverse responses across all six dimensions indicate that philosophical health involves unique configurations of perspectives rather than standardized patterns. The finding that only 12.5% of respondents considered themselves to be "the highest version of my possible selves" suggests widespread recognition of the ongoing challenges of authentic self-development, which is intertwined with the aforementioned principle of autological autonomy.

7.4 The Principle of Epistemic Humility

The fourth principle concerns maintaining appropriate awareness of the limitations of both human and artificial intelligence in philosophical understanding. "Epistemic humility" refers to proper recognition of the boundaries and uncertainties in our knowledge. This principle holds that AI systems should acknowledge the limitations of algorithmic approaches to philosophical understanding rather than trying to answer at any cost – the cost of "hallucinating" at the expense of humans. As Code (1987) argues, epistemic responsibility requires awareness of the limits of knowledge and the social contexts that shape it. In digital philosophical health, AI systems should acknowledge areas of uncertainty, contestation, and mystery rather than presenting artificial certainty. This aligns with Socratic wisdom – knowing what one does not know – and with recognition that most knowledge relies on perspectives rather than imposing universal objectivity (Haraway 1988). The PHI survey findings indirectly indicate the importance of epistemic humility in philosophical health. The substantial minority (33.3%) who viewed the universe as "a sort of mind" alongside those (20.8%) who viewed it as "random" suggest the legitimate plurality of perspectives on fundamental questions. And the diverse responses regarding the sense of the possible indicate varying degrees of epistemic confidence about future potentialities.

These four principles – autological autonomy (or autonomous self-development), hermeneutic integrity, philosophical authenticity, and epistemic humility – provide a possible ethical framework to deal by design with AI integration in philosophical health. They address the unique ethical considerations of using technology to support human

philosophical reflection while avoiding the pitfalls of technological determinism or human displacement.

For instance, the Philai prototype of philosophical health companion, accessible via the Philosophical Health International site, has been designed such that it will often refer to a human counsellor rather than answering questions that might involve an existential risk. For instance, if the human expresses suicidal ideation in the chat, Philai responds: "If you're experiencing thoughts of suicide, please reach out immediately to professional crisis support services or emergency services in your area. Your life has value and there are people who want to help. I am just a chatbot, perhaps it's better to have a direct conversation with a human philosophical counsellor."

8 Concluding Remarks and Future Directions

The integration of AI into philosophical health practice has implications that extend beyond immediate applications to broader questions about the nature and future of philosophical inquiry in the digital age. AI integration challenges traditional boundaries between professional philosophers and broader publics. Philosophical counsellors trained in the SMILE_PH method often observe that individuals without formal philosophical training can with proper guidance engage with profound philosophical questions across all six dimensions of philosophical health. AI-assisted philosophical practice can democratize access to philosophical reflection, potentially realizing a pragmatic enlightenment where philosophical inquiry becomes more widely accessible. AI could help us accelerate a cultural mutation by which humans would rely less on their subcortical structures and more on their philosophical health capacities.

AI integration also raises questions about the relationship between systematic and intuitive modes of philosophical thinking. As McGilchrist (2009) argues, Western intellectual traditions have increasingly privileged analytical, systematic thinking over contextual, embodied understanding. The C.I.P.H.E.R. model (Crealectic Intelligence and Philosophical Health for Eudynamic Realities), with its integration of existential creativity and personal wisdom, suggests a potential rebalancing of these complementary modes of understanding, along forms of cognition that are embodied, extended, enactive and embedded in lifeworlds.

AI integration equally illuminates tensions between universalist and particularist approaches to philosophical health. The diversity of philosophical perspectives revealed in the PHI survey suggests that philosophical health involves singular configurations as well as more universal patterns. Yet AI systems typically seek generalizable patterns across diverse instances. This tension requires ongoing negotiation between recognizing common patterns and respecting individual uniqueness. Critical thinking would insist on the difference between statistical generalization and the idea of a universal concept.

AI integration must also highlight the relationship between philosophical reflection and practical action. The PHI poll finding of a gap between purpose recognition and implementation reflects broader questions about how philosophical insight translates into lived practice. It is not clear if AI assistance can help bridge this gap by supporting ongoing integration of reflection and action, potentially addressing what Schön (1983) identified as the challenge of reflection-in-action. Ideas need to be tested in agonistic or intercreative dialogue with the world, not just before a screen.

Finally, AI's increasing integration into our existences raises questions about the future of philosophical practice itself. As technological capabilities continue to evolve, the boundaries between human and artificial philosophical reflection may become increasingly complex. This evolution demands ongoing examination of what constitutes creative understanding and meaningful philosophical development in a technologically mediated world.

Collective work is needed to rethink the nature, purpose, and possibilities of philosophical inquiry in the digital age. By approaching "anthrobotics", the symbiosis of man and machine, with both critical and intercreative exploration, we may develop approaches that enhance eudaimonic well-being and preserve or even democratically expand the essential qualities that make philosophy a uniquely human endeavor.

An expansion of empirical research on philosophical health across diverse populations could provide richer understanding of how philosophical perspectives vary across cultures, ages, social contexts, and life circumstances. Such research could inform more inclusive and responsive approaches to philosophical health that acknowledge the diversity of human philosophical experience. Conversely, we might be surprised by the emergence of universal structures in planetary thinking.

An optimistic exploration of hybrid human-AI philosophical communities could transcend the individual focus of philosophical counselling to address the social dimensions of philosophical health highlighted in the quantitative findings. Such communities might combine AI assistance with human dialogue to create what Nussbaum (1997) calls "cosmopolitan communities of inquiry" that cross traditional boundaries while fostering genuine philosophical exchange.

Philosophical health is not an abstract ideal but a lived reality with diverse manifestations across individuals and

contexts. The contemporary prevalence of loneliness despite connection, the widespread and perhaps frantic belief in an evolving self, the tension between perceived possibility and false agency – these trends reveal philosophical needs that only creative human thinking can address. The responsibility of thinking for oneself and taking an existentially important decision should never be delegated to another, human or machine.

Funding Open Access funding provided by University of Turku (including Turku University Central Hospital). This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Actions grant agreement No. 101,081,293. The author has no relevant financial or non-financial interests to disclose.

Declarations

Ethics Approval and Consent to Participate No ethical approval was needed as the data was collected anonymously.

Competing Interests The authors has no competing interests to declare that are relevant to the content of this article.

References

Bandura A (1977) Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 84(2):191–215. https://doi.org/10.1037/0033-295X.84.2.191

Berry DM, Fagerjord A (2017) Digital humanities: knowledge and critique in a digital age. Wiley

Boden MA (2004) The creative mind: Myths and mechanisms. Routledge

Borgmann A (1984) Technology and the character of contemporary life: A philosophical inquiry. University of Chicago Press

Brey P (2000) Disclosive computer ethics. ACM SIGCAS Computers Soc 30(4):10–16

Brown G, Chiek Y (2016) Leibniz on compossibility and possible worlds. Springer

Buber M (1970) I and thou. Charles Scribner's Sons

Cacioppo JT, Cacioppo S (2018) Loneliness in the modern age: an evolutionary theory of loneliness (ETL). In: Olson JM (ed) Advances in experimental social psychology. Elsevier Academic, Cambridge, pp 127–197

Carr N (2010) The shallows: what the internet is doing to our brains. W.W. Norton

Code L (1987) Epistemic responsibility. University Press of New England

de Miranda L (2017) On the concept of creal: the politico-ethical horizon of a creative absolute. In: de Assis P, Giudici P (eds) The dark precursor: Deleuze and artistic research. Leuven University

de Miranda L (2020a) Ensemblance: the transnational genealogy of esprit de corps. Edinburgh University

de Miranda L (2020b) Artificial intelligence and philosophical creativity: from analytics to crealectics. Hum Affairs 30(4):597–607. https://doi.org/10.1515/humaff-2020-0053

de Miranda L (2023) Introducing the SMILE_PH method: Sensemaking interviews looking at elements of philosophical health. Methodological Innovations 16(2):163–177. https://doi.org/10.1177/20597991231179336

de Miranda L (2024a) Philosophical health: A practical introduction. Bloomsbury

de Miranda L (2024b) The crealectic method: from creativity to compossibility. Qualitative Inq 1(8). https://doi.org/10.1177/10778004241229065

de Miranda L, Ramamoorthy S, Rovatsos M (2016) We anthrobot: learning from human forms of interaction and esprit de corps to develop more diverse social robotics. In: Seibt J et al (eds) What social robots can and should do. IOS

de Miranda L, Levy R, Divanoglou A (2023) Tapping into the unimpossible: philosophical health in lives with spinal cord injury. J Eval Clin Pract 29(7):1203–1210. https://doi.org/10.1111/jep.13874

Dreyfus HL (1992) What computers still can't do: A critique of artificial reason. MIT Press

Dreyfus HL (2001) On the internet. Routledge

Engelbart DC (1962) Augmenting human intellect: A conceptual framework. Stanford Research Institute

Ess C (2004) Critical thinking and the bible in the age of new media. University Press of America

Floridi L (2014) The fourth revolution: how the infosphere is reshaping human reality. Oxford University Press

Floridi L, Taddeo M (2016) What is data ethics? Philosophical Trans Royal Soc A: Math Phys Eng Sci 374(2083):20160360

Frankfurt HG (1988) The importance of what we care about. Cambridge University Press

Gadamer H-G (1975) Truth and method. Seabury

Gallagher S (2000) Philosophical conceptions of the self: implications for cognitive science. Trends Cogn Sci 4(1):14–21

Gallagher S (2020) Action and interaction. Oxford University Press

Goff P (2019) Galileo's error: Foundations for a new science of consciousness. Pantheon

Hadot P (1995) Philosophy as a way of life: spiritual exercises from Socrates to Foucault. Blackwell

Haraway D (1988) Situated knowledges: the science question in feminism and the privilege of partial perspective. Feminist Stud 14(3):575–599

Heidegger M (1962) Being and time. Harper & Row. (Original work published 1927)

Ihde D (1990) Technology and the lifeworld: from garden to Earth. Indiana University Press

James I, Ardeman-Merten R, Kihlgren A (2014) Ontological security in nursing homes for older persons - person-centred care is the power of balance. Open Nurs J 8:79–87. https://doi.org/10.2174/1874434601408010079

Johnson M (2007) The meaning of the body. University of Chicago Press

Keegan P (2012) Teaching philosophy online. Teach Philos 35(3):277–291

Korsgaard CM (1996) The sources of normativity. Cambridge University Press

Kroger J, Marcia JE (2011) The identity statuses: origins, meanings, and interpretations. In: Schwartz SJ, Luyckx K, Vignoles VL (eds) Handbook of identity theory and research. Springer, pp 31–53
Lara F, Deckers J (2019) Artificial intelligence as a socratic assistant for moral enhancement. Neuroethics 13:275–287
Licklider JCR (1960) Man-computer symbiosis. IRE Trans Hum Factors Electron HFE–1(1):4–11
Manders-Huits N (2011) What values in design? The challenge of incorporating moral values into design. Sci Eng Ethics 17(2):271–287
McGilchrist I (2009) The master and his emissary: the divided brain and the making of the Western world. Yale University Press
Merleau-Ponty M (2012) Phenomenology of perception. Routledge. (Original work published 1945)
Nissenbaum H (2010) Privacy in context: technology, policy, and the integrity of social life. Stanford University Press
Nussbaum MC (1997) Cultivating humanity. Harvard University Press
O'Neill O (2002) Autonomy and trust in bioethics. Cambridge University Press
Pea RD (2004) The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. J Learn Sci 13(3):423–451
Peters MA, Jandrić P (2018) The digital university: A dialogue and manifesto. Peter Lang Publishing
Polanyi M (1966) The Tacit dimension. University of Chicago Press
Ricoeur P (1992) Oneself as another. University of Chicago Press
Russell S (2019) Human compatible: Artificial intelligence and the problem of control. Viking
Sartre JP (2007) Existentialism is a humanism. Yale University Press. (Original work published 1946)
Schön DA (1983) The reflective practitioner: how professionals think in action. Basic Books
Searle JR (2010) Making the social world. Oxford University Press
Stokes PD (2002) Creativity: symbolic equivalence and variation. Eur J High Ability 13(1):95–110
Taylor C (1989) Sources of the self: the making of the modern identity. Harvard University Press
Taylor C (1991) The ethics of authenticity. Harvard University Press
Turkle S (2011) Alone together: why we expect more from technology and less from each other. Basic Books
Tylka TL, Wood-Barcalow NL (2015) What is and what is not positive body image? Conceptual foundations and construct definition. Body Image 14:118–129
van Dijk J (2020) The digital divide. Wiley
Youyou W, Kosinski M, Stillwell D (2015) Computer-based personality judgments are more accurate than those made by humans. Proc Natl Acad Sci 112(4):1036–1040
Zahavi D (2005) Subjectivity and selfhood: investigating the first-person perspective. MIT Press