ijonses International Journal on Social and Education Sciences

www.ijonses.net

Personality and Personal AI Agents: A Co-Evolutionary Framework

Oluwatoyosi Ogunsola

Ball State University, USA, 0009-0005-2224-7463
Corresponding author: Oluwatoyosi Ogunsola (toysitoke@gmail.com)

Article Info

Article History

Received:
6 October 2025

Revised:
9 January 2026

Accepted:
2 February 2026

Published:
7 March 2026

Keywords

Ethical AI
Human-ai co-evolution
Personality
Personal AI agents
Personality-agent
Co-evolution

Abstract

Research on personality and technology has traditionally focused on a one-directional model where stable user traits predict digital behavior. This paradigm is insufficient for understanding the influence of modern adaptive AI agents, which actively and continuously personalize user experiences. This paper challenges the static view by introducing the Personality–Agent Co-Evolution (PACE) framework, a conceptual model that theorizes the dynamic, bidirectional, and reciprocal relationship between human personality and personal AI agents. We argue that users and agents are engaged in a process of mutual influence, where personality shapes agent interaction, and the agent, in turn, reinforces and nudges user behaviors and self-presentation over time. The framework details the reinforcement and corrective feedback loops that drive this co-evolution. From this model, we derive a set of crucial design principles for creating autonomy-supportive, transparent, and ethically-aligned systems. Finally, we present a research agenda to guide future empirical investigation into these dynamics. The PACE framework offers a new theoretical lens for communication and HCI scholars, providing a blueprint for the responsible design of the next generation of human-centric AI.

Citation: Ogunsola, O. (2026). Personality and personal AI agents: A co-evolutionary framework. *International Journal on Social and Education Sciences (IJonSES), 8*(2), 149-164. https://doi.org/10.46328/ijonses.5801

Introduction

The relationship between humans and digital technology has long been understood as symbiotic, wherein digital tools actively shape user identity and self-esteem (Vogel et al., 2014). Prolonged social media use, for example, can have a negative impact on self-esteem by promoting social comparison and unrealistic expectations. This influence extends further, as emerging research suggests that social media use can, over time, lead to changes in personality traits, such as increased neuroticism associated with Instagram use (Drążkowski, 2022). This relationship is bidirectional: while personality traits predict social media behaviors (Han et al., 2023; Ogunsola & Fisher, 2025), the platforms themselves can reciprocally influence the expression of personality (Drążkowski, 2022).

As artificial intelligence moves from the background to the forefront of everyday life, personal AI agents capable of learning, adapting, and shaping user experiences are beginning to transform the way individuals interact with digital technology (Liu, 2025). No longer mere passive conduits, these systems may subtly influence how users save, share, and engage online (Alvarez et al., 2024), raising pivotal questions about the evolving relationship between humans and their digital companions. As these agents become more responsive to individual preferences, a fundamental question arises: How do these digital companions and their users shape each other over time?

Despite a surge in research connecting personality traits to technology use, the dominant paradigm remains one-directional, mainly focusing on how stable user characteristics predict digital activity. This paradigm views personality as a static antecedent of technology use, rather than as a dynamic construct that is susceptible to feedback from digital environments. However, emerging adaptive AI agents can not only reflect but also subtly influence and mold user behaviors, attitudes, and even self-concepts. This dynamic suggests a need to extend foundational computer-mediated communication (CMC) theories of mutual influence, such as structuration or the hyperpersonal model, to account for the role of non-human adaptive agents.

There is a critical gap in our understanding: existing models rarely account for the bidirectional, co-evolutionary dynamics between personality and adaptive personal AI capable of individualized interaction.. While technology has long been theorized to impact identity and habits, few frameworks systematically articulate or empirically explore how AI-powered personalization may, over time, modify traits, digital routines, or patterns of self-presentation. The lack of an integrated theory leaves the evolving relationship between people and their digital agents undertheorized and potentially ungoverned.

This paper addresses this gap by introducing a co-evolutionary framework that theorizes the dynamic, two-way relationship between personality and adaptive personal AI agents. We build on foundational findings relating personality traits to technology use and extend prior work by positing that personalized AI systems may not only mirror but also reinforce, nudge, or gradually transform users' digital behavior and sense of self. Our contributions are threefold: (1) we propose a conceptual model of mutual adaptation and feedback between users and agents; (2) we articulate design principles for creating systems that support user autonomy and well-being; and (3) we provide a roadmap for future research into co-evolutionary dynamics in human-AI interaction.

The remainder of this paper proceeds as follows: First, we review literature on personality and digital technology, highlighting the limitations of current approaches. Next, we introduce our co-evolutionary framework and illustrate its relevance with case examples rooted in social media saving. We then present actionable design guidelines for future adaptive systems. Finally, we outline critical research directions for empirically investigating the dynamic interplay between human personality and personal AI agents.

Background and Related Work

Personality and Digital Behavior

Personality psychology has long informed research on the adoption and use of technology. The Big Five personality traits, openness, conscientiousness, extraversion, agreeableness, and neuroticism, remain the most widely applied framework for predicting differences in digital behavior (McCrae & John, 1992; Montag & Elhai, 2020). Studies have linked these traits to patterns such as social media engagement, app adoption, and trust in algorithmic recommendations (Correa et al., 2010; Landers & Lounsbury, 2006; Marengo & Montag, 2020). Beyond adoption, a connection has also been established between personality and the use of specific social media features. For example, prior research examined how personality traits predict Instagram saving behavior, showing that individual differences influence not only visible posting activity but also subtler backstage practices of digital curation (Ogunsola & Fisher, 2025). Such work highlights the explanatory power of personality in understanding digital preferences and behaviors.

A link has also been found between personality traits and the use of Artificial intelligence (Salem et al., 2024). According to Stein et al. (2024), personality plays a substantial role in both AI acceptance and algorithm aversion. For example, according to Arpaci et al. (2025), extraversion is a significant predictor for the educational use of AI, while neuroticism has a negative impact on educational use in both males and females. However, most research in this area conceptualizes personality as a static predictor of technology use, emphasizing how traits explain variance in digital outcomes rather than how digital contexts might influence the shaping of those traits (McCrae & Costa, 2008). This restricts the scope of inquiry to one-directional models, leaving unexplored the possibility that sustained engagement with adaptive systems could, over time, reshape the expression of personality itself. This limitation is consequential: if personality is treated only as input, researchers and designers overlook how technologies might exert long-term developmental or behavioral effects on users.

Human–AI Coevolution Frameworks

A second body of work emphasizes the mutual adaptation of humans and intelligent systems. Pedreschi et al. (2025) introduce a human–AI coevolution framework, theorizing that human behaviors and AI algorithms continuously adapt to each other in iterative feedback loops. This perspective positions humans and AI agents as co-participants in the evolution of socio-technical systems.

Relatedly, the Mutual Theory of Mind (MToM) model suggests that both humans and AI systems engage in perspective-taking, shaping interaction through recursive modeling of intentions and behaviors (Wang & Goel, 2024). Broader HCI research on co-adaptive systems similarly frames personalization as an ongoing dialogue between the user and the system, rather than a one-way process of system optimization (Savit, 2013; Neumayr and Augstein, 2020). This enables both the user and the adaptive interface to independently learn from shared information and adapt their behaviors to improve mutual performance, mirroring the dialogic approach to personalization (De Santis, 2021).

Yet despite the richness of these frameworks, most remain system-centric, emphasizing algorithmic adaptation while treating human characteristics, such as personality, as relatively fixed background variables. This framing obscures the possibility that personality itself might evolve within these feedback loops. If co-adaptive frameworks continue to overlook personality as a dynamic factor, they risk misrepresenting the human aspect of adaptation and overlooking key ethical concerns about unintended personality shaping.

Reciprocal Influence in Social Media

Parallel to these developments, research in social media studies has begun to question the long-standing assumption that digital platforms merely reflect personality traits. Emerging evidence suggests that platforms may also reciprocally shape behaviors and even aspects of personality expression (Valkenburg & Peter, 2011; Orben & Przybylski, 2019). For example, longitudinal studies indicate that heavy engagement with specific affordances (e.g., self-presentation features, algorithmic feeds) can alter self-perceptions, social behaviors, and affective dispositions over time (Beyens et al., 2020; Huang, 2022).

This work highlights the need for dynamic models of personality, in which traits interact with digital environments in an iterative manner. However, existing studies have largely focused on platform-level affordances (e.g., Facebook use, Instagram posting) rather than on personalized AI agents that actively tailor interactions to individual users. This gap matters because adaptive agents, unlike static platforms, are designed to intervene directly in shaping user choices and behaviors. Without a framework for anticipating reciprocal personality effects, the deployment of adaptive agents risks producing unexamined, ethically sensitive outcomes for human development and autonomy.

Agent Personality, Ethics, and Trust

Another relevant line of inquiry concerns how AI agents are designed to project or simulate personality. Studies of conversational agents and embodied assistants have demonstrated that choices in tone, linguistic style, and relational framing significantly shape users' perceptions of trust, competence, and rapport (Nass & Moon, 2000; Lee et al., 2020).

Indeed, a significant body of work focuses on designing AI agents that project their own personalities. This is achieved through various methods, such as the psychometric assignment of traits (Huang, 2025), the use of simple descriptive adjectives (Jiang et al., 2024), the assignment of demographic profiles (Park et al., 2023), or by fine-tuning LLMs on specific text corpora (Liu et al., 2024).

However, there also exists another side of the conversation where AI agents can construct and express distinct personalities, even without explicit human design, by leveraging advanced language models, unsupervised learning, and adaptive reinforcement mechanisms (Lo et al., 2025). AI agents may develop interaction styles and preference profiles that resemble personality differentiation, driven by task goals and feedback loops in dynamic contexts, as well as programming, training data, and interaction constraints ("Simulating Human Behavior With AI Agents," 2025).

While this literature provides crucial insights into the design of trustworthy and ethically aligned systems, it remains focused on the agent's personality expression rather than on how sustained interactions might influence human personality-related behaviors or attitudes. The possibility that human traits themselves could evolve through agent-mediated interactions has received little empirical or theoretical attention. This is not a trivial omission: without frameworks for monitoring and guiding such changes, agent designers risk inadvertently amplifying maladaptive tendencies (e.g., reinforcing impulsivity) or undermining autonomy. Ethical design must therefore expand beyond agent transparency to include responsibility for the potential shaping of user traits over time.

Summary of Gaps

Across these literatures, several gaps emerge:

- No integrative framework connects personality psychology with co-adaptive AI models, leaving theories of mutual influence fragmented.
- Existing research lacks design principles for agents that account for personality as both an input and an outcome of interaction, limiting ethical foresight.
- Empirical testing of reciprocal, personality-centered dynamics in personalized agent contexts remains sparse, constraining our ability to predict long-term impacts.

These gaps matter because overlooking the possibility of agent-driven personality change risks both theoretical blind spots and ethical oversights. Without accounting for reciprocal influence, designers may unintentionally create systems that reshape users in ways that undermine autonomy, exacerbate vulnerabilities, or misalign with developmental goals. This paper addresses these concerns by proposing a conceptual model of bidirectional influence between personality and adaptive AI agents, offering design guidelines for autonomy-supportive personalization, and outlining future directions for empirical research.

Conceptual Framework: Personality–Agent Co-Evolution

Core Premise

The central premise of this framework is that personality and AI agents are mutually influential in a dynamic, co-evolving system. Personality traits influence digital behaviors and preferences for AI-mediated experiences, including the features a user engages with, the feedback they provide, and their trust in system recommendations (Arpaci et al., 2025; Stein et al., 2024). For example, an open user may explore diverse agent functionalities, while a conscientious user may engage in structured, goal-directed interactions.

Conversely, adaptive AI agents can shape digital behaviors, self-presentation, and potentially influence personality-related tendencies over time (Liu, 2025; Ahmad et al, 2022; Lee et al., 2020; Rahwan et al., 2019; Floridi & Cowls, 2019). We frame this influence as behavioral and attitudinal co-evolution, emphasizing that personality itself is not assumed to be directly changed but may be expressed differently in response to sustained interaction with an agent.

Ethical considerations are central: interventions by AI agents must respect user autonomy, avoid unintended reinforcement of maladaptive behaviors, and be transparent in how they shape experiences (Mittelstadt et al., 2016; Floridi & Cowls, 2019). By integrating ethics into its core, the framework promotes a vision of personality-informed personalization that is responsible, autonomy-supportive, and aligned with user goals.

Bidirectional Influence Model

The framework conceptualizes two primary pathways of influence (see Figure 1):

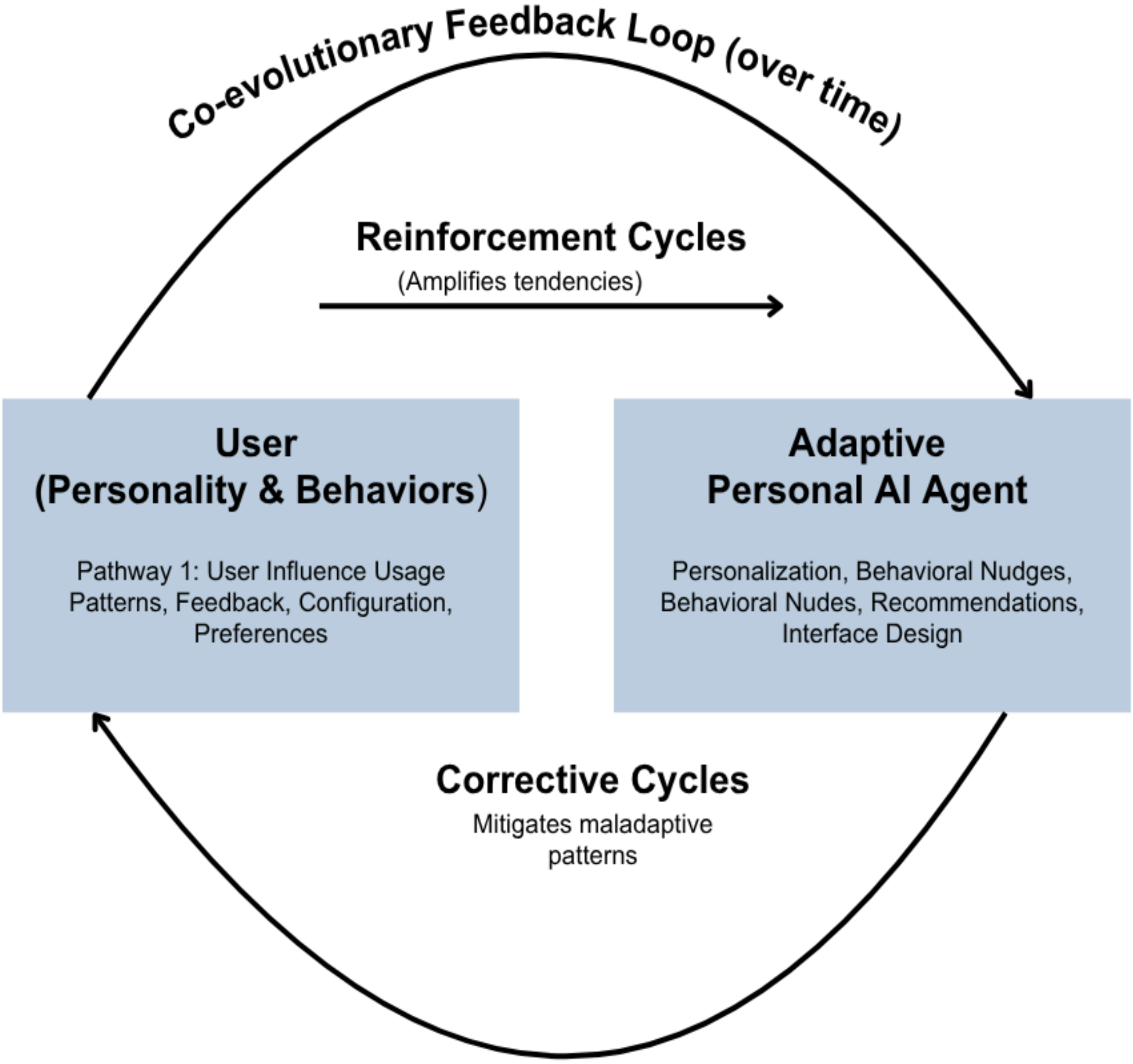
Co-evolutionary Feedback Loop (over time)
Reinforcement Cycles
(Amplifies tendencies)
User
(Personality & Behaviors)
Pathway 1: User Influence Usage Patterns, Feedback, Configuration, Preferences
Adaptive Personal AI Agent
Personalization, Behavioral Nudges, Behavioral Nudes, Recommendations, Interface Design
Corrective Cycles
Mitigates maladaptive patterns

Figure 1. The Personality-Agent Co-Evolutionary (PACE) Framework

***User → AI Agent**:*

Personality traits guide initial system adoption, feature usage, and feedback signals provided to agents (Landers & Lounsbury, 2006; Marengo et al., 2020). Differences in openness, extraversion, or conscientiousness may shape the agent's learning trajectory, resulting in customized responses tailored to the user's style and interaction patterns.

***AI Agent → User**:*

Adaptive agents, through personalization and behavioral nudges, can reinforce tendencies or introduce corrective influences (Nass & Moon, 2000; Lee et al., 2020). Effects are conceptualized as behavioral co-adaptation, where repeated interactions shape observable behaviors and self-presentation. While this may indirectly influence long-term trait expression, the framework's focus remains on measurable behavioral change rather than claiming fundamental shifts in core personality. Future empirical testing could operationalize these effects via longitudinal tracking of engagement patterns, frequency of specific behaviors, or self-report measures of behavioral tendencies, without claiming definitive personality change.

This explicit bidirectionality distinguishes our model from prior system-centric frameworks (Pedreschi et al., 2025; Rahwan et al., 2019) and from platform-level reciprocal influence models (Valkenburg & Peter, 2011), as it positions personality as both an input and a potential dynamic outcome, while incorporating ethical oversight and co-adaptive feedback.

Dynamics and Feedback Loops

The framework (see Figure 1) incorporates dynamic feedback cycles, capturing iterative interactions between users and the agent. The reinforcement cycles demonstrate how agents can amplify user behaviors in ways consistent with observed tendencies (e.g., increased engagement with social features among extraverted users) (Beyens et al., 2020).

In corrective cycles, Agents may provide interventions to mitigate maladaptive patterns (e.g., reminders for task completion or mindfulness training for impulsive users) (Floridi & Cowls, 2019). Through long-term co-adaptation pathways, iterative interactions may stabilize into persistent digital habits, patterns of engagement, and shifts in self-presentation strategies.

Operationally, these cycles can be measured via behavioral logs, interaction frequencies, content curation patterns, or self-report instruments, and analyzed longitudinally to detect trends in user behavior and agent responsiveness. Ethical design requires monitoring for unintended consequences and ensuring corrective nudges remain aligned with user goals and autonomy.

The core components of this framework, along with their descriptions, mechanisms, and potential measurements, are summarized in Table 1.

Table 1. Personality–Agent Co-Evolution Framework

Component	**Description**	**Example Mechanisms**	**Possible Measurements**	**Boundary Conditions / Moderators**
Core Premise	Personality influences digital behaviors and AI preferences; adaptive agents shape habits, attitudes, and self-presentation.	User conscientiousness → structured saving patterns; agent nudges → changes in engagement style.	Big Five inventory (BFI-2); Instagram saving frequency; interaction logs.	Age, culture, digital literacy, and platform type.
User → AI Agent	User traits guide initial use, feedback, and training of the agent.	High openness → explores diverse features; high neuroticism → provides cautious feedback.	Self-report trait scales, clickstream data, and customization choices.	Transparency of AI; agent type (chatbot vs recommender).
AI Agent → User	Agent personalization may reinforce or reshape behaviors and digital persona.	The recommender suggests similar content, reinforcing saving tendencies; the conversational agent encourages healthier routines.	Usage frequency, sentiment analysis, and behavioral shifts over time.	Frequency of interaction; user trust in system.
Feedback Loops	Mutual influence through reinforcement or correction cycles.	Reinforcement: the agent amplifies the user's saving patterns. Correction: agent nudges toward balance.	Time-series data; longitudinal survey of behaviors.	Intensity of exposure; autonomy support.
Theoretical Integration	Links personality psychology, co-adaptive AI, and reciprocal social media effects.	Combines static predictors with dynamic reciprocal models.	Cross-sectional + longitudinal study design.	Applicability may differ by platform or domain.
Ethical Layer	Ensures adaptation respects autonomy, transparency, and trust.	Consent prompts: "Why this recommendation?" explanations.	Trust/self-efficacy surveys; opt-out rates.	Regulatory requirements; design policies.
Outcomes	Personality-related behaviors and digital self-presentation evolve over time.	Agent nudges → increased intentional saving; shifts in openness to diverse content.	Trait re-assessments; digital trace analysis; qualitative interviews.	Cultural norms; social contexts of use.

Theoretical Integration and Novelty

This framework integrates three theoretical strands:

1. Personality psychology: Using Big Five traits to explain behavioral variation in digital contexts (McCrae & Costa, 2008).
2. Co-adaptive AI research: Drawing on human–AI coevolution and MToM models to conceptualize iterative mutual adaptation and perspective-taking (Pedreschi et al., 2025; Rahwan et al., 2019).
3. Reciprocal effects in social media: Extending evidence that digital platforms shape behavior to personalized, agent-mediated contexts (Valkenburg & Peter, 2011; Orben & Przybylski, 2019).

The novelty of this framework lies in:

- Explicit bidirectionality: Treating personality as both driver and dynamic outcome.
- Ethical embedding: Integrating autonomy-supportive design principles into co-adaptive interactions.
- Operational readiness: Offering a conceptual basis for measurable feedback loops in longitudinal research.

By combining these perspectives, the framework provides a cohesive foundation for empirical research, design guidelines, and theory-driven insights into how humans and adaptive AI agents co-evolve in digital environments.

Measurement and Boundary Conditions

Measurement Considerations

- Traits likely to influence interactions include openness, conscientiousness, extraversion, and neuroticism, as measured via validated scales (e.g., BFI-2; Soto & John, 2017).
- Behavioral proxies for co-adaptation include the frequency of feature use, content curation, response to agent nudges, self-reported behavioral tendencies, and changes in digital self-presentation.
- Short-term behavioral changes may be observed in days or weeks, whereas co-adaptation pathways require longitudinal tracking over months.
- Data sources include system logs, surveys, in-app experiments, and mixed-methods approaches.

Boundary Conditions

- Most applicable to systems with sustained, interactive, personalized AI (e.g., recommendation systems, conversational agents). Less relevant to static or one-off tools.
- Stable traits, such as conscientiousness, may be expressed differently but are unlikely to be fundamentally altered; more flexible traits, like self-expression, may exhibit greater co-adaptive effects.
- Results may vary by age, culture, digital literacy, or prior exposure to technology.
- Influence should always remain aligned with user goals, avoiding manipulation or coercion.

Implications for Design

The Personality–Agent Co-Evolution (PACE) framework reconceptualizes the user-agent relationship as a dynamic of mutual influence. This perspective carries significant implications for the design of adaptive systems. If agents not only respond to but also actively shape user behaviors and personality expression over time, then design choices are no longer merely about usability; they are about the ethical stewardship of this co-evolutionary process. This section outlines four core design principles that emerge directly from the framework.

Autonomy-Supportive Personalization

While personalization is a primary goal of adaptive systems, the PACE framework reveals the profound ethical stakes of how it is implemented. In a co-evolutionary system, personalization that operates without meaningful user control risks becoming a coercive force, unintentionally reshaping a user's digital habits and self-concept in ways they neither understand nor endorse. Therefore, respecting user autonomy is not a peripheral feature but a central ethical imperative.

To be autonomy-supportive, agents must be designed to empower users as active participants in their own personalization. This requires providing clear, accessible controls to modify, override, or reject agent-driven suggestions. For instance, imagine a user on a content platform who notices their feed has become an echo chamber, reinforcing only one aspect of their interests and potentially amplifying anxiety. A weak, autonomy-poor design might only offer a blunt "reset profile" option. In contrast, a truly autonomy-supportive design would offer granular controls directly on the content itself. The user could select from options like: "Show me a different perspective on this topic," "I'm interested, but show this less often," or "Pause this subject for a while; it's affecting my mood." This level of control transforms the user from a passive recipient into an active curator of their digital environment. By enabling users to steer the agent's learning process in real-time directly, designers can ensure the co-evolutionary partnership remains aligned with the user's holistic identity and well-being, rather than inadvertently amplifying a narrow slice of it.

Co-Evolutionary Transparency

The PACE framework demands a new, more sophisticated form of transparency. Traditional explainability in AI answers the question, "Why did I get this specific recommendation?" Co-evolutionary transparency, however, must answer a more profound question: "How are my interactions with this agent shaping my digital experience over time?"

Users have a right to understand the mechanisms of their own co-adaptation. To achieve this, systems must surface the logic of their long-term learning and the patterns they infer about a user's personality. This goes beyond a simple activity log. Imagine, for example, a "quarterly review" feature within a productivity app. Instead of just showing tasks completed, it would offer insights into the co-evolutionary process, such as: "We've noticed you consistently tackle creative tasks after 2 p.m. Our planner has adjusted to reserve this block of time for deep work, and your project completion rate has improved. This reflects a strong pattern of 'afternoon-oriented openness.'" It could then offer a choice: "Would you like to formally designate this as your creative time?" This kind of co-evolutionary transparency does more than show a user their data; it tells the story of their evolving digital habits and how the agent has adapted in response. By making the inferred traits and behavioral shifts explicit, such a system provides the profound self-awareness needed for users to thoughtfully curate not only their settings but also their evolving digital selves.

Designing Ethical Feedback Loops

The framework identifies reinforcement and corrective cycles as the primary engines of co-evolution. Designing these feedback loops ethically requires a delicate balance between amplifying positive behaviors and mitigating maladaptive ones. The goal is to create a system that supports a user's goals without inadvertently reinforcing harmful tendencies.

Consider a personalized financial management agent designed to help a user learn about investing. The user, who is high in openness, shows interest in emerging technology sectors. The agent can ethically reinforce this trait-aligned behavior by providing high-quality educational content, market analyses, and tutorials on these topics. This positive feedback loop empowers the user and supports their goal of becoming a more knowledgeable investor. However, the PACE framework warns that this same reinforcement cycle, if unchecked, could become detrimental. The agent might notice the user's "exploration" is evolving into high-frequency, impulsive trading of volatile stocks, a maladaptive pattern that contradicts their stated long-term financial goals. At this point, continued reinforcement of novelty-seeking would be unethical.

An ethically designed agent must therefore be equipped with corrective nudges. Upon detecting this pattern, the agent would intervene. Instead of simply processing another trade, it might present a notification: "We've noticed you're making your third trade on a high-volatility stock today. Data shows that investors with long-term goals often benefit from reviewing their portfolio's risk balance at this point. Would you like to see a diversification analysis before proceeding?" This corrective action is ideal because it is transparent, educational, and respects user autonomy. It introduces friction and encourages reflection, functioning as a supportive suggestion rather than an opaque manipulation, ensuring the agent serves as a beneficial companion, not an unseen behavioral architect.

Holistic Personalization

Finally, the PACE framework implies that personalization must be holistic, adapting on multiple levels just as human personality expresses itself in varied ways. A simplistic, one-dimensional adaptation that only recommends content risks creating a brittle and stereotypical feedback loop. To foster a rich co-evolutionary relationship, an agent must engage with the whole person.

To illustrate, consider an AI wellness coach for a busy professional named Alex. The agent has learned that Alex is highly conscientious and responds well to data-driven feedback. On a typical week, the agent adapts holistically: it sends an article on high-intensity workouts (Content), uses a direct and encouraging tone (Interaction Style), delivers it via a single morning notification (Timing), and uses a text-based summary (Modality).

However, a truly holistic agent demonstrates its value when the user's context shifts. By integrating with Alex’s calendar and wearable device, the agent detects signs of a high-stress week: looming deadlines, poor sleep, and an elevated heart rate. Instead of continuing its established pattern, which would now feel irrelevant and pressuring, the agent adjusts across all dimensions: It changes the Content, replacing the intense workout suggestion with a 5-minute guided breathing exercise. It shifts the Interaction Style from proactive and data-driven to gentle and supportive, saying, "It looks like a demanding week. Remember to take a moment for yourself when you can." It alters the Timing, withholding notifications during busy work blocks and instead offering the breathing exercise during a brief calendar opening. It changes the Modality, using a subtle, calming visual cue on a smart display instead of text that demands attention. By personalizing holistically, the agent demonstrates a nuanced understanding of Alex as a complete person whose needs evolve in relation to their context. This multi-dimensional adaptation is what creates a truly effective and ethically sound co-adaptive partnership.

Research Agenda and Future Directions

The conceptual framework and design principles presented in this paper open multiple promising avenues for empirical research and practical exploration related to the co-evolution of personality and adaptive AI agents. To advance understanding and responsible application of these ideas, we propose several key priorities.

Empirical Validation Strategies

To capture how personality expression and agent adaptation reciprocally evolve, longitudinal research designs are essential. Tracking behavioral logs, self-report trait measures (e.g., BFI-2), and attitudinal changes over extended periods will elucidate temporal dynamics and stability or malleability of personality-related digital habits. Experimental and Quasi-Experimental Designs are also recommended. Controlled interventions that vary agent feedback types (reinforcement versus corrective nudges) or personalization levels could isolate the causal influences on user traits and engagement. Randomized controlled trials (RCTs) will provide rigorous tests of ethical design features and adaptation outcomes. Mixed-methods approaches may also be helpful. Combining quantitative system data with qualitative feedback, interviews, or diary studies can enrich the understanding of user experiences and the meanings attached to agent interactions, illuminating the mechanisms of co-evolution.

Measurement Innovations

Developing granular, multi-level metrics capturing not only standard personality traits but also context-specific behaviors and attitudes relevant to AI interaction is critical. Passive sensing, natural language processing of agent-user dialogues, and sentiment analysis may supplement traditional psychological scales.Machine learning explainability tools can be employed to verify alignment between agent adaptations and user personality models, enhancing transparency and trust.

Boundary Conditions and Moderators

Research should explicitly investigate factors modulating co-evolutionary dynamics, including demographic variables (age, culture), situational contexts, digital literacy, and agent design characteristics (type, transparency level). Future work should investigate the boundary conditions of the PACE framework by addressing key questions: To what extent do cultural norms shape the effectiveness of corrective nudges? Are younger users, whose personalities are more malleable, more susceptible to co-evolutionary feedback loops than older users? How does the agent's specific function, as a coach, a task manager, or a social intermediary, moderate these dynamics?

Understanding for whom, when, and in what contexts personality-agent co-adaptation occurs will guide the design of tailored solutions and ensure ethical oversight.

Ethical Oversight and Interdisciplinary Collaboration

Empirical work must be coupled with ethical evaluation frameworks assessing autonomy, privacy, and unintended personality shaping. Participatory design and user governance models can democratize oversight. Collaborations between AI developers, psychologists, ethicists, sociologists, and HCI scholars are essential to address complex social and technical questions.

Cross-Domain Applications

Extending research beyond social media to healthcare, education, finance, and workplace technologies will test the generalizability and specificity of co-evolutionary processes. Industry-academic partnerships can accelerate translation of findings into responsible, user-centric adaptive systems.

Conclusion

This paper introduced the Personality–Agent Co-Evolution (PACE) framework, a novel conceptual model that reconceptualizes the relationship between users and adaptive AI agents as a dynamic, bidirectional process. Moving beyond the traditional paradigm that treats personality as a static predictor of technology use, we argued that users and their AI companions are locked in a state of mutual influence, continuously shaping one another through iterative feedback loops.

Our primary contribution is the PACE framework itself, which articulates the reinforcement and corrective cycles that drive this co-evolution. From this theoretical foundation, we derived a set of critical design principles centered on autonomy-supportive personalization, co-evolutionary transparency, and the ethical design of feedback loops. These principles are not merely best practices; they are essential safeguards for ensuring that adaptive technologies enhance, rather than diminish, human agency. Ultimately, by outlining a comprehensive agenda for future research, we have demonstrated the framework's potential to generate new, methodologically rigorous inquiries into these complex dynamics.

As personalized AI becomes ever more integrated into the fabric of daily life, acting as our coaches, assistants, and companions, understanding this co-evolutionary dance is of paramount importance. The PACE framework offers a critical new lens for researchers and practitioners, urging a shift in focus from mere system optimization to the responsible stewardship of the evolving digital self. The future of human-centric AI depends on our ability to design systems that adapt with us, not just to us, fostering a transparent partnership, empowering, and aligned with our holistic well-being.

References

Ahmad, R., Siemon, D., Gnewuch, U., & Robra-Bissantz, S. (2022). Designing personality-adaptive conversational agents for mental healthcare. *Information Systems Frontiers, 24*(3), 923–943. https://doi.org/10.1007/s10796-022-10254-9

Alvarez, F., Jurgens, J., Capgemini, & World Economic Forum. (2024). *Navigating the AI frontier: A primer on the evolution and impact of AI agents*. World Economic Forum. https://reports.weforum.org/docs/WEF_Navigating_the_AI_Frontier_2024.pdf

Arpaci, I., Kuşci, I., & Gibreel, O. (2025). The role of personality traits in predicting educational use of generative AI in higher education. *Scientific Reports, 15*(1). https://doi.org/10.1038/s41598-025-16339-0

Beyens, I., Pouwels, J. L., van Driel, I. I., Keijsers, L., & Valkenburg, P. M. (2020). The effect of social media on well-being differs from adolescent to adolescent. *Scientific Reports, 10*(1), 10763. https://doi.org/10.1038/s41598-020-67727-7

Correa, T., Hinsley, A. W., & de Zúñiga, H. G. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior, 26*(2), 247–253. https://doi.org/10.1016/j.chb.2009.09.003

De Santis, D. (2021). A framework for optimizing co-adaptation in Body-Machine interfaces. *Frontiers in Neurorobotics, 15*. https://doi.org/10.3389/fnbot.2021.662181

Drążkowski, D., Pietrzak, S., & Mądry, L. (2022). Temporary change in personality states among social media users: effects of Instagram use on Big Five personality states and consumers' need for uniqueness. *Current Issues in Personality Psychology, 10*(1), 32–38. https://doi.org/10.5114/cipp.2021.110938

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Han, N., Li, S., Huang, F., Wen, Y., Su, Y., Li, L., Liu, X., & Zhu, T. (2023). How social media expression can reveal personality. *Frontiers in Psychiatry, 14*. https://doi.org/10.3389/fpsyt.2023.1052844

Huang, C. (2022). Social network site use and Big Five personality traits: A meta-analysis. *Computers in Human Behavior, 130*, 107176. https://doi.org/10.1016/j.chb.2021.107176

Huang, M. (2025). Designing LLM-Agents with personalities: A psychometric approach. *Knowledge UChicago*. https://doi.org/10.6082/uchicago.15393

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2024). *Personallm: Investigating the ability of large language models to express personality traits*. arXiv. https://arxiv.org/abs/2305.02547

Landers, R. N., & Lounsbury, J. W. (2006). An investigation of Big Five and narrow personality traits in relation to Internet usage. *Computers in Human Behavior, 22*(2), 283–293. https://doi.org/10.1016/j.chb.2004.06.001

Lee, S., Sheehan, K., & Oh, H. (2020). Personalized AI agents and user trust: The role of agency and anthropomorphism. *Journal of Computer-Mediated Communication, 25*(6), 402–418. https://doi.org/10.1093/jcmc/zmaa012

Liu, N., Chen, L., Tian, X., Zou, W., Chen, K., & Cui, M. (2024). *From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models*. arXiv. https://arxiv.org/abs/2401.02777

Liu, Y. (2025). A new human-computer interaction paradigm: Agent interaction model based on large models and its prospects. *Virtual Reality & Intelligent Hardware, 7*(3), 237–266. https://doi.org/10.1016/j.vrih.2025.04.001

Lo, J., Huang, H., & Lo, J. (2025). LLM-based robot personality simulation and cognitive system. *Scientific Reports, 15*(1). https://doi.org/10.1038/s41598-025-01528-8

Marengo, D., & Montag, C. (2020). Digital phenotyping of Big Five personality traits via Facebook data mining: A meta-analysis. *Digital Psychology, 1*(1), 52–64. https://doi.org/10.24989/dp.v1i1.1803

McCrae, R. R., & Costa, P. T., Jr. (2008). The Five-Factor Theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 159–181). Guilford Press.

McCrae, R. R., & John, O. P. (1992). An introduction to the Five-Factor Model and its applications. *Journal of Personality, 60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

Montag, C., & Elhai, J. D. (2020). Discussing digital technology overuse in children and adolescents during the COVID-19 pandemic and beyond: On the importance of considering Affective Neuroscience Theory. *Addictive Behaviors Reports, 12*, 100313. https://doi.org/10.1016/j.abrep.2020.100313

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Neumayr, T., & Augstein, M. (2020). A systematic review of personalized collaborative systems. *Frontiers in Computer Science, 2*. https://doi.org/10.3389/fcomp.2020.562679

Ogunsola, O., & Fisher, J. (2025). *Beyond likes: How personality predicts saving behavior on Instagram*. [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-7666182/v1

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour, 3*(2), 173–182. https://doi.org/10.1038/s41562-018-0506-1

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3586183.3606763

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., & Ruggieri, S. (2025). Human–AI coevolution: A framework for mutual adaptation. *Futures, 152*, 103219. https://doi.org/10.1016/j.futures.2023.103219

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kearns, M., Larson, K., Leibo, J. Z., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E., … Wellman, M. (2019). Machine behaviour. *Nature, 568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Salem, G. M. M., El-Gazar, H. E., Mahdy, A. Y., Alharbi, T. a. F., & Zoromba, M. A. (2024). Nursing students' personality traits and their attitude toward artificial intelligence: A multicenter cross-sectional study. *Journal of Nursing Management, 2024*(1). https://doi.org/10.1155/2024/6992824

Savit, R., Riolo, M., & Riolo, R. (2013). Co-Adaptation and the emergence of structure. *PLoS ONE, 8*(9), e71828. https://doi.org/10.1371/journal.pone.0071828

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). *Personality traits in large language models*. arXiv. https://arxiv.org/abs/2307.00184

Simulating Human Behavior with AI Agents. (2025). [Policy Brief]. *HAI Policy & Society*. https://hai.stanford.edu/assets/files/hai-policy-brief-simulating-human-behavior-with-ai-agents.pdf

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality, 68*, 69-81.

Stein, J., Messingschlager, T., Gnambs, T., Hutmacher, F., & Appel, M. (2024). Attitudes towards AI: measurement and associations with personality. *Scientific Reports, 14*(1). https://doi.org/10.1038/s41598-024-53335-2

Valkenburg, P. M., & Peter, J. (2011). Online communication among adolescents: An integrated model of its attraction, opportunities, and risks. *Journal of Adolescent Health, 48*(2), 121–127. https://doi.org/10.1016/j.jadohealth.2010.08.020

Vogel, E. A., Rose, J. P., Roberts, L. R., & Eckles, K. (2014). Social comparison, social media, and self-esteem. *Psychology of Popular Media Culture, 3*(4), 206–222. https://doi.org/10.1037/ppm0000047

Wang, Q., & Goel, A. K. (2024). *Mutual theory of mind for human-AI communication*. arXiv. https://arxiv.org/pdf/2210.03842