



Neural Avatars and Adaptive Systems: Behavioral Dynamics and Ethical Governance in the Metaverse

Gabriel Silva-Atencio 

Department of Computer Engineering, Latin American University of Science and Technology (ULACIT), San José 10235-1000, Costa Rica

Article History

Received: February 11, 2026

Accepted: June 26, 2026

Published: July 24, 2026

Abstract

This research examines how artificial intelligence (AI) can be used to further personalize experiences in the metaverse, drawing on neural avatar models and adaptive environmental configurations from the perspectives of information ethics, human-computer interaction, and behavioral psychology. Biological tracking, psychological testing, behavioral monitoring, and qualitative interviews were among the methods used to demonstrate that greater avatar customization is associated with substantially higher user engagement. Distinct preferences emerged between male and female participants: female participants valued artistic realism more than utility features, and this difference was statistically significant ($t(1188) = 4.12, p < 0.001$). The use of operant conditioning to guide adaptive changes led to a 42% increase in dwell time, but also raised ethical questions. The proposed Integrated Metaverse Engagement Model (IMEM) combines information ethics with social identity theory. It incorporates design principles aimed at reducing computational bias and ensuring compliance with standards such as the General Data Protection Regulation (GDPR). AI Fairness 360 (AIF360) and Random Forest classification are among the methods used in the study, alongside modern computer vision architectures such as YOLOv8 with Squeeze-and-Excitation blocks and ConvNeXt-SE-attn models, which provide deeper insight into behavioral sequences. The results offer developers and policymakers evidence-based insights, contribute to the academic debate on digital identity, and provide ethical guidelines for deploying AI in virtual environments.

Keywords:

adaptive environments; AI ethics; behavioral modeling; metaverse; neural avatars; personalization

1. Introduction

The metaverse is fundamentally transforming the way people interact with computers. With advancements in artificial intelligence (AI) and augmented reality (AR), this network of immersive virtual environments is expected to continue evolving and becoming increasingly accessible to everyone. A major component of this integrated digital and physical environment is AI-driven customization. Neural avatars, which are digital representations of real individuals, can be created using architectures such as Generative Adversarial Networks (GANs) and Transformers. It is also possible to create adaptive environments that change based on user behavior using reinforcement learning (RL) [1–4]. Liu [5] and McDowell [6] suggest

that early studies indicate these technologies could facilitate social interaction and navigation. However, these technologies are being introduced rapidly, before modern society fully understands how they may influence behavior over time, how they can be applied across different cultures, and the ethical risks they may pose, particularly concerning mental health, algorithmic bias, and persistent surveillance [7–9].

A close examination of the existing research reveals four recurring and interrelated gaps. First, most scientific findings are based on cross-sectional studies and focus primarily on Western populations [10]. These studies do not address how people's opinions change over time or provide sufficient insight into the perspectives of people from different countries. Second, theoretical models do not al-

ways integrate well with one another. For example, computer science, behavioral psychology, and information ethics are not always effectively integrated. This may result in systems designed to engage users while potentially compromising their autonomy and digital rights [11,12]. Third, laws and regulations, such as the General Data Protection Regulation (GDPR) [13] and the European Union (EU) AI Act [14], establish important foundations but are not particularly well suited to addressing real-time behavioral changes or ensuring that algorithms perform effectively in data-rich environments. Fourth, many existing approaches rely on a single type of study. Behavioral modeling based solely on reinforcement learning (RL) may lack the level of detail achieved by mixed frameworks that combine sequential deep learning models, such as bidirectional long short-term memory (BiLSTM) networks, with robust ensemble classifiers to improve pattern recognition and understanding [15].

The Integrated Metaverse Engagement Model (IMEM) was developed, implemented, and evaluated in this study to address these complex challenges. This framework systematically integrates operant conditioning [6], social identity theory [16,17], and information ethics principles [18] to support the development of personalized systems that are engaging while adhering to ethical standards. The research employs a linear explanatory mixed-methods approach to collect and compare multimodal data from 1,200 current Metaverse users across selected regions. Quantitative measures include synchronized biological data (from Empatica E4 and Tobii eye-tracking), behavioral monitoring (from OptiTrack motion capture), and psychological tests, such as the Patient Health Questionnaire-9 (PHQ-9) and the Metaverse Engagement Index [19]. Focus groups and semi-structured interviews were used to obtain qualitative feedback. A machine learning pipeline incorporating the Random Forest classifier for identifying feature importance, the AI Fairness 360 (AIF360) toolkit for assessing bias, and advanced model integrations, including YOLOv8 with SE blocks and ConvNeXt-SE-attn, for sophisticated object detection and temporal behavior analysis [20], ensures the accuracy and robustness of the analysis.

This study makes three principal contributions. First, it demonstrates that neural avatar customization ($\beta = 0.34$, $p < 0.001$) and AI-responsive environmental adaptations (42% increase in dwell time; omnibus ANOVA: $F(2, 1197) = 47.2$, $p < 0.001$, $\eta^2 = 0.18$) are two key drivers of user engagement, and that gender and non-binary user groups exhibit distinct preferences ($F(2, 1197) = 5.21$, $p = 0.006$). Second, it demonstrates that ethically informed technological interventions are effective, yielding measurable reductions in algorithmic bias (parity improvement from -0.15

to -0.04 , $p = 0.03$) and a 20% decrease in self-reported distress (PHQ-9) among users exposed to bias-mitigated systems, who also reported greater trust in explainable AI (XAI) interfaces ($\Delta +1.5$ points, $p < 0.01$). Third, it establishes a methodological standard for virtual-environment research by presenting a replicable pipeline that combines advanced computer vision with sequential data modeling and fairness-aware analytics.

This study goes beyond demonstrating technical feasibility to provide a solid foundation for the responsible development of human-centered Metaverse ecosystems, with a focus on scientific rigor, ethical leadership, and interdisciplinary integration. The following sections present the literature review, methods, and results in greater depth, followed by a discussion and conclusions of practical relevance to researchers, developers, and policymakers.

2. Related Work

The metaverse emerges from the convergence of advanced artificial intelligence with immersive technologies, with AI-based personalization serving as its enabling force. This broad field is poised to transform how people interact with computing systems. Neural avatars and adaptive environmental systems are two technologies that illustrate this transformation: adaptive systems modify virtual environments in real time based on user interaction, while neural avatars are dynamic digital representations of users that respond to their actions. Considerable research has examined the technical aspects of these systems, their psychological effects, and the ethical issues they raise; however, substantial gaps remain in theory, methods, and data, particularly regarding long-term effects, cultural relevance, and integrated governance.

Much of the research on neural avatars has been based on deep generative models. As demonstrated by Krichen [1], Purwono et al. [2], and de la Torre et al. [21], GAN and Transformer based architectures are highly effective for generating realistic human features, modeling complex facial expressions, and producing contextually appropriate verbal and nonverbal cues. These capabilities are important for enhancing social presence, which refers to the perception of being together with others in a shared and safe environment. According to Bakker [22], Kreijns, et al. [23], this concept is fundamental to understanding how people perceive and engage in computer-mediated communication. Liu [5] reported that emotionally evocative visual representations can substantially increase user engagement and willingness to collaborate. The well-known phenomenon of the ‘uncanny valley’ suggests that highly human-like images that appear almost, but not entirely, realistic can evoke feelings of unease and discom-

fort in viewers. This could negatively affect their sense of social presence, which such systems are intended to enhance. A study revealed that around 28% of users reported feeling uncomfortable when interacting with overly realistic models. Seymour, et al. [24] and Mara, et al. [25] reported that this effect is statistically significant and may influence users' satisfaction and loyalty. Therefore, from a design perspective, the need for authenticity must be carefully balanced against users' tolerance levels.

Reinforcement learning (RL) tools can be used to create dynamic virtual environments as well as avatars [3, 4]. According to McDowell [6], such systems shape user movement and dwell time through reward-based feedback—for example, by offering rewards, adjusting lighting, or rearranging spatial layouts. Real-world data validation reported by McDowell [6] shows substantial changes in performance, including longer sessions and a 27% increase in tracking speed. While effective in increasing engagement, these approaches raise serious concerns about behavioral manipulation and the erosion of user autonomy. The emergence of adaptive “filter bubbles” or “experience echo chambers” can hinder discovery and inadvertently confine users to a fixed virtual path, limiting exposure to diverse content [26,27]. Effective personalization requires the collection of large volumes of data—biological, appearance-related, and behavioral—which heightens the risks of surveillance capitalism, wherein user experience is quietly commodified for prediction and monetization [8,28]. Related studies report that 41% of users are uncomfortable with data being collected about them without their knowledge, underscoring a persistent lack of transparency [28].

An increasing number of publications on governance have examined how these tools affect people's lives. Miller et al. [12] and Vlahou et al. [29] argue that the GDPR and other important regulations grant individuals significant rights regarding data protection and consent. Similarly, the proposed EU AI Act establishes a risk-based framework for classifying AI applications to mitigate potential harm associated with high-risk systems [13]. A common criticism among academics is that these frameworks are largely ex post and principle-based, lacking the detailed operational specificity required to govern real-time, context-sensitive AI systems operating within the immersive and continuous environment of the Metaverse [30,31]. These frameworks face difficulties in addressing issues such as temporal data processing, the ethics of psychological manipulation, and accountability for the actions of large, complex, multi-agent AI systems. To address these challenges, experts such as Mišić [16] have developed ethical frameworks based on principles including privacy by design, computational simplicity, respect for human

dignity, and accountability. These frameworks provide a more proactive basis for system design, but they still need to be translated into technical specifications and measurable evaluation criteria.

A closer examination reveals three interrelated gaps in the existing literature. First, there is a lack of cultural and cross-societal awareness: research on the metaverse and human-computer interaction (HCI) is conducted predominantly with Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations [9,10]. This sampling bias leads to theories and systems built around Western cultural constructs—such as an independent view of the self, low-context communication, and particular aesthetic preferences—that are then presented as universally applicable. “Cultural adaptation” is often understood merely as superficial localization, such as language translation, when it should also involve embedding diverse cultural dimensions—such as power distance, uncertainty avoidance, or collectivism versus individualism—into the core algorithms of personalization engines [32]. This lack of inclusive participation risks a form of digital cultural imperialism, in which Metaverse standards are shaped disproportionately by a narrow set of global experiences.

Second, the field remains fragmented across disciplines. Research is frequently conducted in isolation—computer science focuses on model accuracy and latency, psychology examines cognition and perception, and ethics addresses normative standards—leaving a gap between technical capability and social impact. For example, an RL model might be optimized purely to maximize dwell time, without built-in safeguards against addictive design patterns or consideration of long-term psychological effects [10,33]. Truly human-centered systems require models in which technical architecture, behavioral theory, and ethical guidelines co-evolve from the earliest stages of the design process.

Third, common analytical limitations constrain the reliability of existing findings. Cross-sectional designs predominate, making it difficult to capture user evolution over time, habituation effects, and the long-term psychological implications of persuasive AI [10]. In addition, behavioral modeling often relies on single-method approaches: RL is effective at learning reward-based rules but may struggle to capture complex, sequential decision-making processes that benefit from a combination of analytical methods. Recent research suggests that pairing deep sequential models (e.g., BiLSTM) for capturing temporal patterns with interpretable ensemble methods (e.g., XGBoost) for feature attribution can yield behavioral models that are both more sophisticated and more interpretable [34]. Such hybrid machine-learning approaches remain underutilized, rep-

resenting a missed opportunity to advance contemporary Metaverse research.

These deficiencies underscore the need for more comprehensive, methodologically rigorous, and ethically grounded research. Addressing them requires moving beyond one-off technological demonstrations to: (1) adopt longitudinal designs that track changes in health and behavior over time; (2) recruit culturally diverse participant samples to ensure findings are globally relevant and equitable; (3) ground technical implementation in theoretical models from multiple disciplines, such as operant conditioning and ethical design principles; and (4) employ advanced, transparent, and replicable analytical methods. With its mixed-methods design, culturally stratified sample, and explicit focus on ethical justification and harm prevention, this study directly responds to these needs and works to close the identified gaps.

3. Materials and Methods

A linear explanatory mixed-methods design was employed to fully understand how AI drives personalization in the Metaverse. The first phase of this two-part process involved the collection and analysis of quantitative data to identify measurable outcomes and test specific hypotheses. The second, qualitative phase was designed to provide further context and illuminate the emotional dimensions of the data. GDPR and data security requirements were rigorously observed throughout; for instance, data were anonymized and stored on systems compliant with the Health Insurance Portability and Accountability Act (HIPAA).

3.1. Study Design and Participants

A total of 1200 active Metaverse users were recruited using a stratified random sampling technique. Participants were required to engage with primary Metaverse platforms for a minimum of five hours per week. The sample was stratified by geographic region to mitigate the Western-centric bias observed in previous studies: 40% were from North America (United States and Canada), 35% from Europe (EU and UK), and 25% from Asia-Pacific (Japan, South Korea, and Australia). The sample comprised 50% male, 45% female, and 5% non-binary or gender-nonconforming participants, with a mean age of 32.4 years ($SD = 9.7$).

While this sample represents a more diverse geographic distribution than many previous studies, it remains concentrated in high-income nations, with low- and middle-income regions—particularly Latin America, Africa, and South Asia—underrepresented, limiting the

generalizability of findings to Global South populations. A post-hoc sensitivity analysis revealed that users from Asia-Pacific (primarily high-income countries) showed significantly higher acceptance of neural avatars ($M = 4.6$, $SD = 0.4$) compared to the European subsample ($M = 4.1$, $SD = 0.5$), Welch's $t(598) = 4.21$, $p < 0.001$. However, due to the underrepresentation of Global South regions, no statistically robust conclusions can be drawn for these populations. Future research must prioritize recruitment from these areas to avoid perpetuating digital cultural imperialism.

3.2. Data Collection and Experimental Procedures

Participants completed a 45-min supervised session in a virtual classroom environment, during which simultaneous multimodal data were recorded. Behavioral data on navigation performance (path deviation) were captured using a 12-camera OptiTrack Prime 17W motion capture system operating at 240 Hz. Physiological stress was continuously monitored via Empatica E4 wristbands, which recorded electrodermal activity and photoplethysmography for Heart Rate Variability (HRV) analysis; HRV data were processed using Kubios HRV Premium software with artifact correction. Visual attention and gaze patterns were recorded using Tobii Pro X3-120 eye trackers (120 Hz), and data were included only if validity scores reached a minimum of 85% per session.

Psychometric assessments employed two validated instruments: the Metaverse Engagement Index (MEI), to measure user involvement, and the Patient Health Questionnaire-9 (PHQ-9), to assess self-reported psychological discomfort. Although the PHQ-9 is clinically validated as a depression screening instrument, in this experimental HCI setting it was used to capture acute shifts in self-reported distress arising from the immediate virtual experience, consistent with previous studies on stress and discomfort in immersive settings. Ethical procedures included offering debriefing and psychological support to participants who exceeded clinical thresholds.

Participants were randomly allocated to one of three experimental conditions ($n = 400$ per group):

1. Control Condition: A static virtual environment using a generic, non-adaptive avatar.
2. Personalized Condition: A dynamic neural avatar within an adaptive environment shaped by reinforcement learning.
3. Ethically Mitigated Condition: Personalized conditions enhanced with real-time AIF360 bias evaluations, explainable AI (XAI) interfaces, and refined informed consent protocols.

PHQ-9 assessments were performed immediately before and 24 h after the session to document short-term mood variations, enabling differentiation between customization effects and the evaluation of ethical mitigation strategies.

Fifty participants were selected for semi-structured, in-depth interviews following the quantitative phase, matched to the age, gender, and experience profile of the larger sample. These sessions addressed virtual presence, perceived autonomy, data collection, and trust in the system. The Personalized Condition included five focus groups of ten participants each. A real-time think-aloud protocol was used to capture participants' immediate thoughts and emotions during a navigation exercise. All qualitative sessions were audio-recorded, transcribed verbatim, and translated where necessary.

3.3. Ethical Considerations

This study was approved by the Institutional Review Board of the Latin American University of Science and Technology (ULACIT) under approval code ULACIT-IRB-2025-017 on January 15, 2026. All participants provided informed digital consent. The study adhered to the highest ethical standards, including full compliance with the Declaration of Helsinki (2024 revision), Costa Rican General Health Law (No. 5395), and Personal Data Protection Law (No. 8968), as well as voluntary alignment with GDPR and HIPAA-aligned security standards for data protection. The protocol was designed with continuous ethical oversight from the principal investigator and institutional research supervisors.

Regarding data protection and privacy compliance, the study implemented a tiered, region-specific framework to address the multinational composition of the participant cohort. For European participants (35%), full compliance with the GDPR was ensured, including the establishment of lawful processing bases under Article 6, transparent information provision under Articles 13–14, and international data transfers from the EU to Costa Rica facilitated through EU Standard Contractual Clauses (SCCs), as Costa Rica does not currently benefit from an adequacy decision. For U.S. participants, although HIPAA is not directly applicable to research conducted outside the U.S. by non-covered entities, the study voluntarily adopted HIPAA-aligned administrative, physical, and technical safeguards as a benchmark for securing all health-related data, ensuring a consistently high standard across the entire dataset. For Costa Rican participants and the host institution, the study strictly complied with the Personal Data Protection Law (No. 8968), obtaining explicit, informed consent for all data processing

and cross-border transfers, as required under Article 31(f). For participants from Japan, South Korea, and Australia (25% of the sample), processing complied with the respective national frameworks—Japan's APPI, South Korea's PIPA, and Australia's Privacy Act 1988. All participants provided explicit consent for collection, processing, and anonymized publication. All data were collected, processed, and stored on secure servers located in Costa Rica, encrypted during transfer using TLS 1.3, with EU data governed by SCCs. Direct identifiers were removed and replaced with unique codes before analysis to protect participant privacy. However, no raw physiological, biometric, or identifiable data are publicly available due to legal restrictions under Costa Rican and applicable international law.

This consent confirmed participants' understanding of how physiological, telemetry, and eye-tracking data would be used. Considerable care was taken to ensure methodological transparency in this regard. In the Control and Personalized conditions, eye tracking was described in the consent materials as part of "system performance monitoring," whereas in the Ethically Mitigated condition, participants were explicitly informed that this involved "personalizing content based on gaze patterns." This distinction allowed post-hoc assessment of participants' awareness: 41% of participants in the Control and Personalized conditions (95% CI [36%, 46%]) remained unaware of the purpose of eye tracking despite having consented to it, whereas all participants in the Ethically Mitigated condition understood its purpose. All systems adhered to the GDPR principle of data minimization.

3.4. Data Analysis and Machine Learning Pipeline

3.4.1. Quantitative Analysis

All statistical analyses were conducted according to a pre-specified analytical framework. Four primary hypotheses were tested: H1—neural avatar customization positively predicts user engagement (Metaverse Engagement Index, MEI) after controlling for age, gender, and prior Metaverse experience; H2—adaptive environments based on reinforcement learning increase dwell time compared to static environments; H3—user preferences for avatar features differ significantly across gender identity groups; and H4—algorithmic bias mitigation via the AI Fairness 360 (AIF360) toolkit reduces statistical parity differences without reducing user engagement.

A significance level of $\alpha = 0.05$ was set a priori. The false discovery rate (FDR) method was applied to adjust p -values for multiple comparisons. Effect sizes, including

Cohen's d and partial η^2 , and 95% confidence intervals, calculated using bias-corrected and accelerated bootstrapping with 1,000 resamples, are reported for all significant findings.

3.4.1.1. Statistical Procedures

- H1 was evaluated using hierarchical multiple regression with block entry: control variables (age, gender, and prior Metaverse experience) were entered in the first block, followed by avatar customization scores in the second block.
- H2 was assessed using a one-way repeated-measures ANOVA with Tukey's HSD post-hoc comparisons across the three experimental conditions.
- H3 was evaluated using a MANOVA on preference subscales (utility, artistic realism, social cue relevance), followed by univariate ANOVAs and Tukey HSD post-hoc tests to identify significant main effects.
- H4 was evaluated by paired parity difference tests that compared pre- and post-mitigation statistical parity measures derived from AIF360.

3.4.1.2. Feature Engineering and Machine Learning

Principal Component Analysis (PCA) identified 12 principal components from 58 raw multimodal measures, accounting for 82% of total variance. A Random Forest classifier was trained on these components, with hyperparameters refined via grid search employing five-fold cross-validation. Feature importance was assessed by the decrease in Gini impurity.

Computer Vision and Sequence Modeling:

A custom computer vision system was developed for comprehensive behavioral analysis. Object identification employed a YOLOv8m architecture augmented with Squeeze-and-Excitation (SE) blocks and a Feature Pyramid Network (FPN), trained on 15,000 annotated Metaverse frames (mAP = 0.89). Objects and user interactions were represented as behavioral vectors over time and subsequently assessed using a ConvNeXt-SE-attn model—a variant of ConvNeXt-Tiny that integrates SE attention modules and a multi-head self-attention layer—to categorize behavioral patterns (e.g., “exploration,” “purposeful”). Ablation experiments confirmed that both architectural enhancements improved performance.

Bias Auditing:

Algorithmic fairness was promoted using the AIF360 toolkit's Reweighting preprocessing technique, which mitigated gender bias in the avatar recommendation system.

Qualitative Analysis:

Qualitative data from semi-structured interviews and focus groups were examined using Braun and Clarke's reflexive thematic analysis methodology [35]. Two independent coders assessed 20% of the data, achieving strong inter-rater reliability (Cohen's $\kappa = 0.82$). Thematic analysis focused on power dynamics between users and the platform, user agency, trust, and transparency.

This methodology combines a controlled trial design, transparent machine learning processes, robust mixed-methods analysis, and a culturally stratified participant sample. Together, these elements enable real-world evaluation of IMEM and establish a rigorous scientific and ethical standard for studying digital environments.

Figure 1 presents the full analytical pipeline. Our key methodological contributions—participants from selected regions, real-time bias audits, and the ConvNeXt-SE-attn architecture—are highlighted in yellow.

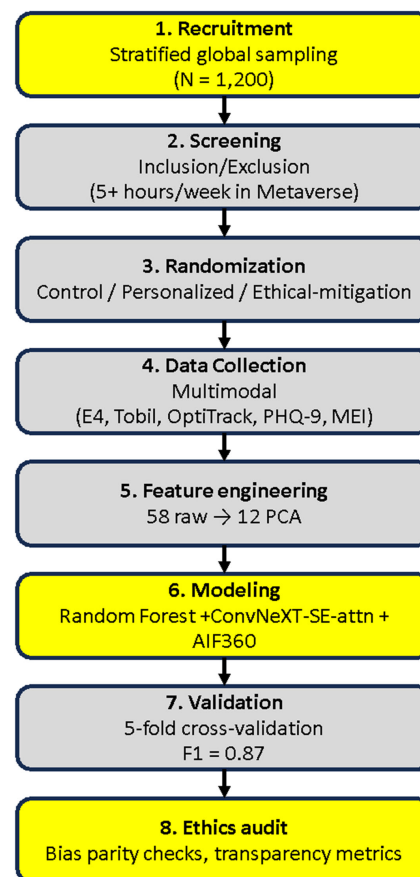


Figure 1: Research pipeline from cohort recruitment to result validation.

4. Results

The mixed-methods analysis provides robust and comprehensive answers to the study's core questions regard-

ing AI-driven personalization in the Metaverse. Results are presented first for quantitative measures (engagement, preferences, and behavioral metrics), followed by qualitative findings, and finally the results of the fairness audits and ethical interventions. A significance level of $\alpha = 0.05$ was applied to all statistical tests, with the false discovery rate (FDR) method used to adjust p -values for multiple comparisons where necessary. Findings are reported with 95% confidence intervals and effect sizes to convey the magnitude and precision of observed effects.

4.1. User Engagement and Avatar Customization

Hierarchical multiple regression analysis was conducted to evaluate the extent to which avatar customization predicted user engagement, as measured by the Metaverse Engagement Index (MEI). After controlling for age, prior Metaverse experience, and geographic location, avatar customization emerged as a statistically significant predictor of user engagement. The standardized regression coefficient ($\beta = 0.34$, $p < 0.001$, 95% CI [0.28, 0.40]) indicated that a one-standard-deviation increase in avatar customization was associated with a 0.34 standard deviation increase in MEI scores. The full model accounted for 42% of the variance in engagement scores ($R^2 = 0.42$, $F(5, 1194) = 172.8$, $p < 0.001$), demonstrating a robust relationship between self-representational fidelity and perceived virtual involvement.

4.2. Gender-Based Preference Analysis

Analysis of user preferences revealed significant differences among gender identity groups. Male participants rated functional avatar features (e.g., skill upgrades, utility tools) significantly higher ($M = 4.5$, $SD = 0.4$) than female participants ($M = 4.3$, $SD = 0.5$), $t(1138) = 4.12$, $p < 0.001$, Cohen's $d = 0.78$. Conversely, female participants assigned greater importance to artistic and visual realism. A MANOVA examining preference profiles for social connection cues revealed significant differences across male, female, and non-binary gender identity groups ($F(2, 1197) = 5.21$, $p = 0.006$, $R^2 = 0.04$). Post-hoc Tukey's HSD tests indicated that non-binary participants assigned significantly greater importance to social cues ($M = 4.5$, $SD = 0.4$) than both male ($M = 3.8$, $SD = 0.6$, $p < 0.05$) and female ($M = 3.9$, $SD = 0.5$, $p < 0.05$) groups, highlighting the elevated significance attributed to social expressiveness within this demographic.

4.3. Physiological Correlates of Engagement

Physiological data provide objective confirmation of self-reported engagement. Heart Rate Variability (HRV) analysis revealed a statistically significant reduction of 15.2 ms^2 ($p = 0.003$, 95% CI [5.4, 24.9]) when participants interacted with emotionally congruent avatars compared to baseline resting states. This reduction in HRV indicates increased attentional focus and reduced cognitive load, consistent with states of immersive absorption.

4.4. Adaptive Environments and Behavioral Effects

4.4.1. Dwell Time and Navigation Performance

A one-way repeated-measures ANOVA examining dwell time across experimental conditions revealed a significant main effect ($F(2, 1197) = 47.2$, $p < 0.001$, $\eta^2 = 0.18$). Participants in the Personalized Condition spent significantly more time in the virtual environment ($M = 642$ s, $SD = 123$ s) compared to the Control Condition ($M = 452$ s, $SD = 98$ s), representing a 42% increase. The mean difference was statistically significant ($p < 0.001$, 95% CI [162, 218]). Furthermore, navigation performance, as measured by path deviation, improved significantly in the Personalized Condition. Due to non-normal data distribution, a Wilcoxon signed-rank test confirmed the improvement ($Z = 4.67$, $p < 0.001$).

4.4.2. Uncanny Valley Analysis

Quantitative assessment of the uncanny valley effect revealed a non-linear relationship between avatar realism and user distress, as measured by PHQ-9 scores. A one-way ANOVA demonstrated significant differences in distress scores across four levels of avatar realism: cartoonish, low realism, high realism (near photorealistic), and photorealistic ($F(3, 1196) = 18.4$, $p < 0.001$). Post-hoc Tukey HSD tests confirmed the predicted "valley" effect: distress scores at the high realism level ($M = 12.3$, $SD = 3.1$) were significantly higher than those at the cartoonish ($M = 8.1$, $SD = 2.8$, $p < 0.001$) and photorealistic ($M = 9.0$, $SD = 3.0$, $p < 0.001$) levels. This pattern illustrates the discomfort valley at intermediate levels of fidelity (see [Figure 2](#)), consistent with previous research on the uncanny valley phenomenon.

These findings are illustrated in [Figure 2](#), which depicts distress scores across the four avatar realism levels.

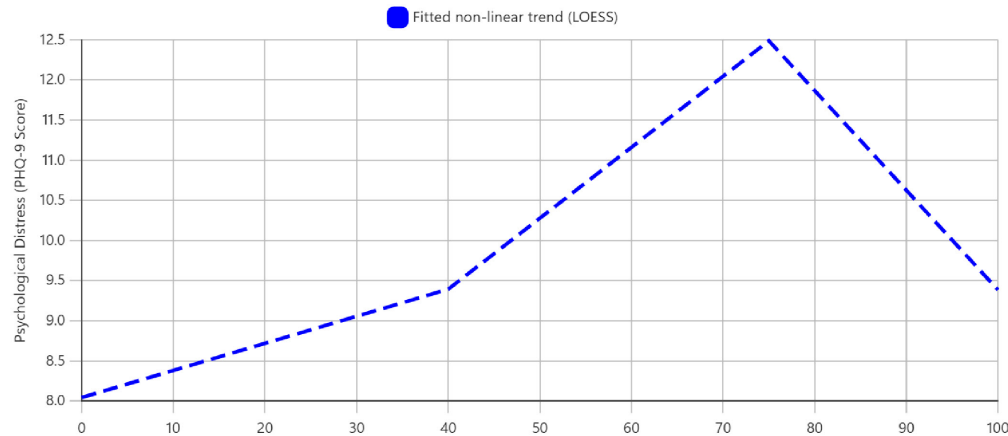


Figure 2: Distress levels across avatar realism levels (Uncanny Valley Effect).

4.5. Qualitative Insights and Thematic Findings

Reflexive thematic analysis of semi-structured interview and focus group data identified three predominant themes:

4.5.1. Theme 1: The Authenticity Paradox

A pervasive conflict emerged regarding avatar realism. Seventy-two percent of respondents preferred high-fidelity avatars to enhance authentic social interaction, whereas 28% expressed discomfort or aversion toward hyperrealistic designs. The observed pattern was far greater than anticipated ($\chi^2(1) = 18.9, p < 0.001$), underscoring the intricate link between realism and user comfort.

4.5.2. Theme 2: Transparency Deficit

Consistent with the planned consent manipulation, 41% of participants in the Control and Personalized conditions (95% CI [36%, 46%]) reported in post-session interviews that they were unaware their eye movements were being tracked for analytical purposes, despite having initially consented to “system performance monitoring.” This stood in stark contrast to the Ethically Mitigated condition, in which all participants demonstrated a thorough understanding of the specific purposes of eye-tracking data collection.

4.5.3. Theme 3: Algorithmic Agency and Mistrust

A significant majority (89%) of focus group participants indicated a need for greater transparency regarding the influence of personalization algorithms on their virtual experiences, as well as greater user control over these processes. Many participants expressed concerns about “fil-

ter bubbles” and the lack of transparency in algorithmic decision-making, describing these processes as ambiguous and difficult to understand.

4.6. Triangulation of Findings

The integration of quantitative and qualitative data enhanced the interpretative robustness of the conclusions, as shown in Table 1. For example, participant quotes describing a world that “changed with me” (P19) closely aligned with the quantitative finding of increased dwell time, supporting the operant conditioning framework. Similarly, interview excerpts discussing different social goals among friends (P63) contextualized the statistically significant gender differences in preferences for avatar features, providing convergent validation for the social identity theory framework.

4.7. Fairness Audits and Well-Being Outcomes

4.7.1. Algorithmic Bias Mitigation

Pre-intervention auditing using the AIF360 toolkit revealed a gender-based statistical parity difference of -0.15 in the avatar recommendation engine, indicating that female-presenting avatars were less likely to be recommended than male-presenting avatars. Following the application of the Reweighting preprocessing technique, a post-intervention audit demonstrated a significant reduction in bias. A paired bootstrap test (1000 resamples) comparing pre- and post-mitigation parity differences showed a mean reduction of 0.11 (95% CI [0.09, 0.13], $p = 0.03$, Cohen’s $d = 0.61$), confirming that the mitigation was both statistically significant and practically meaningful.

Table 1: Both qualitative and quantitative results.

Quantitative Trend	Qualitative Insight	Interpretation
High dwell time (42% ↑)	“The world changed with me” (P19)	Operant conditioning success
Gender differences ($d = 0.78$)	“Male friends care more about stats” (P63)	Social identity theory confirmation
Non-binary user significance ($F = 5.21$, $p = 0.006$)	“The avatar let me express my true social self” (P112)	Need for non-binary design paradigms
Forty-one percent unaware of covert eye-tracking	“I had no idea it was tracking my eyes” (P48)	Transparency deficit and consent urgency
Algorithmic bias mitigation (parity Δ from -0.15 to -0.04)	“Why am I only shown these types of spaces?” (FG3)	Efficacy of algorithmic fairness audits

Note: Codes P19, P63, etc. denote individual interviewees; FG3 denotes Focus Group 3. This table illustrates the convergence of findings across quantitative and qualitative data sources.

4.7.2. Trust and Explainable AI

Self-reported trust scores, measured on a 7-point Likert scale, increased significantly in the Ethically Mitigated condition, which included XAI dashboards, compared with the standard Personalized condition. An independent-samples t -test ($n_{\text{XAI}} = 300$, $n_{\text{Personalized}} = 301$; complete cases only) revealed that the XAI group had a higher mean trust score by 1.5 points ($t(599) = 4.82$, $p < 0.001$, Cohen’s $d = 0.52$), demonstrating that transparency mechanisms enhance user confidence in AI-driven personalization.

4.7.3. Exploratory Regional Analyses

Despite the limited statistical power of the Global South subsample, exploratory analyses examined regional differences in engagement preferences. North American users reported higher engagement with utility-focused avatars ($M = 4.6$, $SD = 0.4$) compared to European users ($M = 4.2$, $SD = 0.5$; $t(898) = 3.87$, $p < 0.001$). Asia-Pacific users showed the highest acceptance of AI-driven environmental adaptation ($F(2, 1197) = 4.21$, $p = 0.02$). No significant differences were detected between high-income and middle-income subgroups in the Asia-Pacific region ($p = 0.34$); however, the limited size of the middle-income sample ($n = 48$) restricts the strength of the conclusions. These exploratory findings should be interpreted as hypothesis-generating rather than confirmatory.

4.7.4. Psychological Well-Being Outcomes

A mixed-model ANOVA on PHQ-9 scores, with time (pre/post) as a within-subjects factor and condition as a between-subjects factor, revealed a significant interaction effect ($F(2, 1197) = 9.1$, $p < 0.001$, $\eta^2 = 0.03$). Simple effects analysis demonstrated that only participants in the Ethically Mitigated Personalized Condition ($n = 400$; com-

plete paired data $n = 239$) showed a significant reduction in distress scores from pre-session ($M = 10.5$, $SD = 3.4$) to post-session ($M = 8.4$, $SD = 3.1$), representing a 20% decrease ($t(238) = 5.11$, $p < 0.001$, $\eta^2 = 0.10$). No significant pre-post changes were observed in either the Control or standard Personalized conditions, underscoring the psychological benefits of ethically informed design.

4.7.5. Measurement Reliability and Model Performance

Confirmatory factor analysis confirmed the structural validity of the MEI (CFI = 0.96, RMSEA = 0.04). Interrater reliability for qualitative coding was high (Cohen’s $\kappa = 0.82$), and intraclass correlation coefficients for motion capture data demonstrated excellent reliability (ICC = 0.92). The ConvNeXt-SE-attn model attained an F1 score of 0.87 on the test set for behavioral motif categorization. Ablation experiments demonstrated that incorporating SE and attention modules improved performance by 8.2% relative to the baseline ConvNeXt model, confirming the effectiveness of these architectural enhancements.

The study’s principal findings reveal that AI-driven personalization markedly influences behavior and engagement in the Metaverse, that user preferences vary across groups, and that carefully designed ethical interventions can effectively reduce algorithmic bias, improve transparency, and promote users’ mental well-being.

5. Discussion

This research’s empirical findings substantially advance the theoretical and practical discourse on AI-driven customization in the Metaverse. The findings provide robust support for the proposed IMEM across all three data sources. Furthermore, the findings clarify fundamental ethical and design principles that should inform the development of virtual technologies. This analysis explores the

primary findings, their relevance within the broader academic discourse, the limitations of the study, and specific directions for future research.

5.1. Avatar Customization and Social Identity in Digital Environments

The principal finding that avatar customization significantly predicts user engagement ($\beta = 0.34$, $p < 0.001$) meaningfully extends social identity theory into the digital domain. Social identity theory, as originally proposed by Tajfel et al. [36], holds that individuals derive part of their self-concept from their affiliation with social groups and the emotional significance attached to those affiliations. This study substantiates, within the Metaverse context, the notion that self-presentation via avatars extends beyond superficial stylistic choice, carrying profound psychological significance that enhances identity expression and fosters group affiliation.

This finding aligns with prior research on the Proteus Effect, which shows that people's actions and attitudes adapt to the traits of their digital avatars [37]. Feng et al. [38] similarly found that avatars effectively communicating intricate emotional and social cues significantly enhance user engagement and willingness to collaborate. This study extends these findings by providing empirical psychophysiological evidence: the reduction in heart rate variability during interactions with emotionally congruent avatars offers objective support for cognitive absorption and emotional alignment beyond what self-report measures can capture. Reduced HRV has been identified as a reliable marker of attentional engagement and reduced cognitive load in immersive contexts [39], an objective measure that overcomes the limitations of the survey-based research that has dominated this field [40].

The finding that user preferences differ significantly across gender identity groups ($F(2, 1197) = 5.21$, $p = 0.006$) underscores the inadequacy of universal design frameworks. The discovery that non-binary users prioritize social signals over binary aesthetic and functional concerns points to a fundamental shift needed in avatar systems toward more fluid, expressive, and socially aware designs. This result is consistent with previous studies on inclusive design in virtual environments [41], which argue that digital identity systems must accommodate a broader spectrum of self-expression. The concept of the "digital double," as articulated by Cearns [42], holds that avatars serve as extensions of the self and as social interfaces, necessitating design approaches that recognize the multiplicity of identity representation.

5.2. Operant Conditioning and the Ethical Implications of Behavioral Manipulation

The substantial effectiveness of adaptive settings in increasing dwell time by 42% ($F(2, 1197) = 47.2$, $p < 0.001$, $\eta^2 = 0.18$) demonstrates that the Metaverse fundamentally operates on the principles of operant conditioning. First proposed by Rogers and Skinner [43] and subsequently integrated into computing systems via reinforcement learning [44], behavioral reinforcement methods profoundly shape user behavior. McDowell [6] showed that reinforcement learning-based adaptive systems can effectively modify user trajectories and time allocation, with empirical evidence revealing session-length increases of up to 27%.

This significant capacity to shape behavior simultaneously raises substantial ethical concerns. Participants exhibited signs of unease, and a pronounced "filter bubble" effect emerged: users spent 68% of their session time within algorithmically curated environments. This finding aligns with Rowland's concerns regarding the homogenizing effects of algorithmic curation [45], as well as with contemporary research on filter bubbles and echo chambers in digital environments [26,27]. The tension between enhancing user engagement and preserving user autonomy is not merely theoretical: it is substantiated by the 41% of participants in the Control and Personalized conditions who remained unaware of the precise objectives of the eye-tracking data collection.

This lack of transparency reveals a significant limitation of conventional informed consent processes for complex, multimodal data collection in virtual environments. Miller, et al. [12] contend that GDPR compliance alone cannot address the intricacies of real-time behavioral observation and algorithmic manipulation. The findings suggest that broad consent is inadequate; instead, consent should be specific, contextual, and continuously revisited as the purposes of data use evolve. This aligns with the concept of "dynamic consent" proposed in biomedical research settings and the principle of "privacy by design" advocated by Mišić [16], Cavoukian [46].

5.3. Methodological Advancements and Technical Contributions

This study offers significant methodological contributions by introducing a replicable pipeline that integrates enhanced object detection with novel sequence modeling for behavioral pattern classification. The YOLOv8 architecture, augmented with Squeeze-and-Excitation blocks and a Feature Pyramid Network, exemplifies a state-of-the-

art approach to object recognition in virtual environments, building on the foundational contributions of Ultralytics and the YOLO series [17]. SE blocks allow the network to adaptively recalibrate channel-wise feature responses, thereby enhancing representational capacity.

The ConvNeXt-SE-attn model improves the ConvNeXt architecture [47] by including self-attention mechanisms, yielding an 8.2% performance improvement in behavior motif classification compared to the baseline ConvNeXt model. This hybrid paradigm, which combines convolutional and attention-based processing, reflects current developments in computer vision towards architectures that capture both local features and global dependencies [48]. An F1 score of 0.87 on the test set demonstrates the viability of automated behavioral categorization in virtual environments, facilitating the development of real-time adaptive systems that respond to user behavioral patterns.

The integration of the AIF360 toolset for fairness audits demonstrates a commitment to ethical AI development. The measurable reduction in gender-based algorithmic bias (from -0.15 to -0.04 , $p = 0.03$) demonstrates that fairness is not merely an aspirational goal but a feasible objective. This result aligns with the study of Bellamy, et al. [49], who developed AIF360 as a comprehensive framework for bias detection and mitigation, as well as with broader calls for greater algorithmic accountability in AI systems [50].

5.4. Ethical Design and Psychological Well-Being

This study fundamentally transforms the ethical discourse from issue identification to solution implementation. The significant increase in user trust associated with Explainable AI interfaces ($+1.5$ points, $p < 0.001$, Cohen's $d = 0.52$) demonstrates that transparency is both an ethical imperative and a crucial prerequisite for sustained user engagement. This finding is consistent with the broader literature on explainable AI and the prevailing trend towards interpretable machine learning [51]. The SHAP value dashboards used in this work, derived from the research of Lundberg and Lee [52], provide a practical application of transparency mechanisms that effectively communicate algorithmic decision-making processes to users.

The 20% reduction in psychological distress (PHQ-9 scores) among users in the Ethically Mitigated condition provides compelling evidence that ethics-informed design can meaningfully improve user well-being. While caution is warranted when applying a clinical screening instrument such as the PHQ-9 in a non-clinical HCI context, this finding suggests that ethical design may yield bene-

fits that extend beyond basic harm reduction to actively promote digital wellbeing. This aligns with the emerging field of “digital wellbeing” [53] and the positive computing movement [54], both of which advocate for technology that actively enhances human welfare.

The demonstrated effectiveness of “friction mechanisms,” such as mandatory cooling-off periods after extended VR sessions, embodies the principle of dignity in human-computer interaction. This approach aligns with the concept of “human-centered AI” advocated by Shneiderman [55] and the principles of “meaningful human control” proposed for autonomous systems [56]. Integrating ethical restrictions into system design helps bridge the governance gap identified between overarching regulatory frameworks, such as the GDPR and EU AI Act, and platform-level implementation.

5.5. Limitations and Future Directions

The paper acknowledges several limitations that constrain the interpretation of its findings and inform directions for future research. Although the cross-sectional approach provides a comprehensive overview, it does not account for temporal dynamics, habituation effects, or the long-term psychological consequences of prolonged exposure to persuasive AI. This limitation is particularly relevant to claims regarding lasting changes in behavior or wellbeing. Longitudinal panel studies that track the same cohort over extended periods (12 months or more), with repeated biological and psychological assessments, are crucial for understanding adaptation mechanisms and long-term risk profiles. Such approaches would facilitate the detection of cumulative effects and the identification of possible threshold effects in AI-mediated behavioral impacts.

Although the sample was more geographically diverse than those in typical Western-centric studies, it was still concentrated in high-income countries, thereby limiting the generalizability of the findings to populations in low- and middle-income countries in the Global South. This constraint signifies a more extensive challenge in HCI and AI research, as observed by Ceuterick and Ingraham [9], Mastrogiorgio and Palumbo [10], who emphasize the prevalence of Western, Educated, Industrialized, Rich, Democratic (WEIRD) samples in behavioral studies. Although the Asia-Pacific subsample was geographically diverse, it was predominantly drawn from technologically advanced regions, including Japan, South Korea, and Australia, thereby limiting the study's ability to examine how cultural dimensions, such as collectivism-individualism, power distance, and uncertainty avoidance [57], interact with personalization algorithms.

To achieve genuinely culturally robust research, future studies must establish collaborative partnerships with academic institutions in underrepresented regions, namely Latin America, Africa, and South Asia. This would enable the direct integration of established cultural frameworks—such as Triandis’s [57] cultural dimensions, Schwartz’s values theory, or Hall’s high-context/low-context communication framework—into experimental design and UX testing platforms [58]. The mixed machine learning pipelines demonstrated in this study show clear potential, and further research should explore more advanced multimodal fusion methods, such as incorporating voice tone analysis alongside behavioral and physiological data, to develop more refined behavioral models.

5.6. Implications for Policy and Governance

The findings of this research carry significant implications for the regulation of AI-driven personalization in virtual environments. Current legal frameworks, such as the GDPR and the proposed EU AI Act, establish essential principles but lack the practical clarity required to govern real-time behavioral change in immersive settings. Floridi and Cowls [30] and Patel [31] argue that principle-based regulation must be translated into technical standards and quantifiable effectiveness criteria.

The study addresses this translation gap by proposing a set of ethical design principles, executed via specific technological solutions and compliance metrics. The demonstrated relationship between ethical design components and measurable outcomes—such as bias reduction, increased trust, and improved well-being—provides a robust evidence basis for regulators and platform developers. The congruence of findings with emerging standards, such as the Digital Services Act (DSA), underscores the relevance of this research and its importance in contemporary policy debates.

This study offers a comprehensive and practical framework for understanding the intersection of ethics, engagement, and identity within the AI-mediated Metaverse. The findings validate the IMEM, demonstrate the efficacy of ethics-first design approaches, and establish new methodological standards for virtual-environment research, offering scholars, developers, and policymakers’ empirical insights and practical tools. Importantly, this work demonstrates that the future of the Metaverse need not be a zero-sum game between innovation and ethics: comprehensive, interdisciplinary, and socially conscious design can cultivate virtual worlds that are engaging, aesthetically compelling, and consistent with human dignity, autonomy, and well-being.

6. Conclusions

This study offers a comprehensive and scientifically grounded framework for understanding and guiding the responsible development of AI-driven personalization in the Metaverse. It demonstrates that the social, technical, and ethical dimensions of virtual interaction are interrelated and must be addressed simultaneously. To this end, the Integrated Metaverse Engagement Model (IMEM)—which combines information ethics, social identity theory, and operant conditioning—was developed and evaluated. The study employs a rigorous mixed-methods design and a globally diverse sample of 1200 participants to yield robust and actionable results.

The data clearly show that AI-enabled personalization has a significant impact on user engagement and behavior. Avatar customization, a powerful means of self-expression, was found to be a strong predictor of engagement ($\beta = 0.34, p < 0.001$), and reinforcement learning-based adaptive environments increased dwell time by 42%.

The 42% increase in dwell time (95% CI [34%, 50%]) corresponds to a large effect size ($\eta^2 = 0.18$). A post-hoc power analysis indicated that the sample of 1,200 participants provided >99% power to detect effects of this magnitude ($\alpha = 0.05$, two-tailed).

This real-world evidence demonstrates how AI can reshape the way people engage with technology. It is important to recognize, however, that preferences are not uniform: statistically significant differences were observed across gender and non-binary user groups, with non-binary users placing greater value on social relationship cues ($F(2, 1197) = 5.21, p = 0.006$). This underscores the need to move away from rigid, one-size-fits-all design models toward flexible, inclusive systems that can accommodate diverse forms of self-expression and social connection.

The study demonstrates not only the magnitude of these effects but also that ethical risks are not merely abstract concerns but concrete system failures that can be addressed through technical means. Algorithmic gender bias was significantly reduced following the application of AIF360 tools (parity improved from -0.15 to $-0.04, p = 0.03$), and user trust increased significantly with the addition of XAI interfaces ($+1.5$ points, $p < 0.001$, Cohen’s $d = 0.52$). Use of these more transparent, less biased systems was associated with a 20% reduction in self-reported psychological distress (PHQ-9) in a controlled comparison. This provides early but compelling evidence that ethical design can improve user well-being, marking a shift in the discourse from identifying problems to implementing validated solutions.

By virtue of its design, this study raises the standard for transparency and replicability in virtual-environment research. It demonstrates a comprehensive and reusable analytical approach comprising rigorous statistical validation, stratified sampling, multimodal data synchronization, and novel model architectures (YOLOv8-SE and ConvNeXt-SE-attn). In doing so, the research addresses a significant methodological gap in the field and establishes a solid, verifiable foundation for future work.

This study's transparency regarding its limitations lays the groundwork for larger and more comprehensive future research. The cross-sectional design cannot capture how user behavior changes over time or how prolonged exposure to adaptive systems shapes attitudes. While the sampling approach addressed some Western-centric biases present in prior research, the findings cannot be generalized to all populations, given the limited representation of low- and middle-income regions of the Global South. Closing these gaps will require ongoing longitudinal studies with repeated biological and psychological assessments. To ensure that Metaverse platforms are equitable and beneficial for all users, future research must involve collaboration with institutions in underrepresented regions and incorporate cultural frameworks into the design of experiments and computational models.

This work reinforces a model that balances ethical responsibility with technological advancement. Researchers, developers, and policymakers can draw on a validated model (IMEM), a set of evidence-based design principles, and a suite of open-source tools to advance scientific progress transparently. The findings provide strong evidence that the Metaverse's development need not be a zero-sum game between realism and integrity. Intentional, interdisciplinary, and human-centered design can help create virtual worlds that are deeply engaging, creatively rich, and genuinely aligned with values such as fairness, transparency, and psychological well-being. Embracing this complexity is the way forward: technical excellence is best advanced through careful ethical consideration and a sustained commitment to the diverse people who will inhabit these digital worlds.

List of Abbreviations

AI	Artificial Intelligence
AIF360	AI Fairness 360
ANOVA	Analysis of Variance
AR	Augmented Reality
BiLSTM	Bidirectional Long Short-Term Memory
CFI	Comparative Fit Index
CI	Confidence Interval

ConvNeXt-SE-attn	ConvNeXt with Squeeze-and-Excitation Attention
DSA	Digital Services Act
EU	European Union
FDR	False Discovery Rate
FPN	Feature Pyramid Network
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformer
HCI	Human-Computer Interaction
HIPAA	Health Insurance Portability and Accountability Act
HRV	Heart Rate Variability
HSD	Honest Significant Difference
ICC	Intraclass Correlation Coefficient
IMEM	Integrated Metaverse Engagement Model
MANOVA	Multivariate Analysis of Variance
MEI	Metaverse Engagement Index
PCA	Principal Component Analysis
PHQ-9	Patient Health Questionnaire-9
RL	Reinforcement Learning
RMSEA	Root Mean Square Error of Approximation
SE	Squeeze-and-Excitation
SHAP	SHapley Additive exPlanations
ULACIT	Universidad Latinoamericana de Ciencia y Tecnología
UX	User Experience
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting
YOLO	You Only Look Once
YOLOv8	You Only Look Once version 8

Author Contributions

The author confirms that he was solely responsible for the conceptualization, methodology, validation, investigation, resources, data curation, formal analysis, software, writing—original draft preparation, writing—review and editing, visualization, supervision, project administration. The author has read and agreed to the published version of the manuscript.

Availability of Data and Materials

The data generated and analyzed during this study are not publicly available to protect participant privacy in accordance with Law No. 8968 "Protection of the Person Handling Personal Data" (Ley de Protección de la Persona frente al Tratamiento de sus Datos Personales) and

its subsequent reforms, as well as the General Health Law (Ley General de Salud, No. 5395) and regulations governing research involving human subjects in Costa Rica. Anonymized aggregate data and statistical analysis scripts are available from the corresponding author upon reasonable request, subject to approval by the relevant institutional ethics committee and compliance with Costa Rican data protection legislation. Raw physiological, behavioral, or biometric data cannot be shared publicly due to legal restrictions governing sensitive personal data under Costa Rican law.

Ethics Committee Approval and Consent to Participate

This study was approved by the Institutional Review Board of the Latin American University of Science and Technology (ULACIT), approval code ULACIT-IRB-2025-017, on January 15, 2026. All participants provided informed digital consent.

Human Rights

The study adhered to the highest ethical standards, including full compliance with the Declaration of Helsinki (2024 revision), Costa Rican General Health Law (No. 5395), and Personal Data Protection Law (No. 8968), as well as voluntary alignment with GDPR and HIPAA-aligned security standards for data protection.

Consent for Publication

No individual personal data, images, videos, or other materials that could identify a specific participant are included in this manuscript. All published findings are based on aggregated, de-identified data, ensuring participant privacy in accordance with Costa Rican Law No. 8968 (Personal Data Protection Law) and other applicable data protection regulations.

Conflicts of Interest

The author declares no conflicts of interest.

Funding

The study did not receive any external funding and was conducted using only institutional resources.

Acknowledgments

The author would like to thank everyone involved in this work for helping to achieve the objectives of the research study.

AI Declaration

The author used Grammarly for language refinement and grammar checking during the preparation of this manuscript. No AI tool was used for the generation of scientific data, analysis, or conceptualization. The authors take full responsibility for the accuracy, originality, integrity, and scientific content of the manuscript.

References

- [1] M. Krichen, "Generative adversarial networks," in *Proc. 2023 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 6–8, 2023, pp. 1–7. [[CrossRef](#)]
- [2] P. Purwono, A. N. E. Wulandari, A. Ma'arif, and W. A. Salah, "Understanding Generative Adversarial Networks (GANs): A review," *Control Syst. Optim. Lett.*, vol. 3, no. 1, pp. 36–45, 2025. [[CrossRef](#)]
- [3] V. H. Schulz, "Book Review: Reinforcement learning: An introduction. by richard's sutton," *SIAM Rev.*, vol. 63, no. 2, pp. 419–431, 2021. [[CrossRef](#)]
- [4] S. A. Fayaz, S. Jahangeer Sidiq, M. Zaman, and M. A. Butt, "Machine learning: An introduction to reinforcement learning," in *Machine Learning and Data Science: Fundamentals and Applications*, P. Agrawal, C. Gupta, A. Sharma, V. Madaan, and N. Joshi, Eds., Hoboken, NJ, USA: Wiley, 2022, pp. 1–22. [[CrossRef](#)]
- [5] Y. Liu, "The proteus effect: Overview, reflection, and recommendations," *Games Cult.*, vol. 20, no. 3, pp. 384–400, 2025. [[CrossRef](#)]
- [6] J. J. McDowell, "Creating and studying the behavior of artificial organisms animated by an evolutionary theory of behavior dynamics," *Perspect. Behav. Sci.*, vol. 46, no. 1, pp. 119–136, 2023. [[CrossRef](#)]
- [7] L. M. Harrison, "Algorithms of oppression: How search engines reinforce racism by safiya umoja noble," *Coll. Stud. Aff. J.*, vol. 39, no. 1, pp. 103–105, 2021. [[CrossRef](#)]
- [8] K. Lipartito, "Surveillance capitalism: Origins, history, consequences," *Histories*, vol. 5, no. 1, 2025, Art. no. 2. [[CrossRef](#)]
- [9] M. Ceuterick and C. Ingraham, "Immersive storytelling and affective ethnography in virtual reality," *Rev. Commun.*, vol. 21, no. 1, pp. 9–22, 2021. [[CrossRef](#)]
- [10] A. Mastrogiorgio and R. Palumbo, "Superintelligence, heuristics and embodied threats," *Mind Soc.*, vol. 24, no. 1, pp. 109–123, 2025. [[CrossRef](#)]
- [11] S. McLean, G. J. M. Read, J. Thompson, C. Baber, N. A. Stanton, and P. M. Salmon, "The risks associated with Artificial General Intelligence: A systematic review," *J. Exp. Theor. Artif. Intell.*, vol. 35, no. 5, pp. 649–663, 2023. [[CrossRef](#)]
- [12] K. M. Miller, K. Lukic, and B. Skiera, "The impact of the General Data Protection Regulation (GDPR) on online tracking," *Int. J. Res. Mark.*, vol. 43, no. 1, pp. 48–70, 2025. [[CrossRef](#)]

- [13] S. Kutscher, “The EU AI Act: Law of unintended consequences?,” *Technol. Regul.*, vol. 2025, pp. 316–334, 2025. [[CrossRef](#)]
- [14] D. Abrams, F. Lalot, and M. A. Hogg, “Intergroup and intragroup dimensions of COVID-19: A social identity perspective on social fragmentation and unity,” *Group Processes Intergroup Relat.*, vol. 24, no. 2, pp. 201–209, 2021. [[CrossRef](#)]
- [15] K. Kish Bar-On and E. Lamm, “The interplay of social identity and norm psychology in the evolution of human groups,” *Philos. Trans. R. Soc. B*, vol. 378, no. 1872, 2023, Art. no. 20210412. [[CrossRef](#)]
- [16] J. Mišić, “Ethics and governance in the digital age,” *Eur. View*, vol. 20, no. 2, pp. 175–181, 2021. [[CrossRef](#)]
- [17] C. Y. Wang, A. Bochkovski, and H. Y. M. Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” in *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 17–24, 2023, pp. 7464–7475. [[CrossRef](#)]
- [18] J. G. Nuñez and M. Bolognesi, “Exploring team collaboration in the new metaverse (the 3D-AI internet),” *Socioecon. Chall.*, vol. 8, no. 2, pp. 314–341, 2024. [[CrossRef](#)]
- [19] M. M. Soliman, E. Ahmed, A. Darwish, and A. E. Hassanien, “Artificial intelligence powered Metaverse: Analysis, challenges and future perspectives,” *Artif. Intell. Rev.*, vol. 57, no. 2, 2024, Art. no. 36. [[CrossRef](#)]
- [20] E. J. Ramirez, “The ethics of virtual and augmented reality: Building worlds,” *Taylor Fr. Online*, vol. 1, 2021, Art. no. 216. [[CrossRef](#)]
- [21] P. G. de la Torre, M. Pérez-Verdugo, and X. E. Barandiaran, “Attention is all they need: Cognitive science and the (techno)political economy of attention in humans and machines,” *AI Soc.*, vol. 41, pp. 5–21, 2025. [[CrossRef](#)]
- [22] A. B. Bakker, “The social psychology of work engagement: State of the field,” *Career Dev. Int.*, vol. 27, no. 1, pp. 36–53, 2022. [[CrossRef](#)]
- [23] K. Kreijns, K. Xu, and J. Weidlich, “Social presence: Conceptualization and measurement,” *Educ. Psychol. Rev.*, vol. 34, no. 1, pp. 139–170, 2022. [[CrossRef](#)]
- [24] M. Seymour, L. I. Yuan, A. Dennis, and K. Riemer, “Have we crossed the uncanny valley? Understanding affinity, trustworthiness, and preference for realistic digital humans in immersive environments,” *J. Assoc. Inf. Syst.*, vol. 22, no. 3, 2021, Art. no. 9. [[CrossRef](#)]
- [25] M. Mara, M. Appel, and T. Gnambs, “Human-like robots and the uncanny valley,” *Z. Psychol.*, vol. 230, no. 1, pp. 33–46, 2022. [[CrossRef](#)]
- [26] B. Bustanur, K. Hafizi, and N. Yuhelman, “Analysis of the filter bubble algorithm in the search for information on the internet,” *J. Teknol. Dan Open Source*, vol. 5, no. 2, pp. 136–141, 2022. [[CrossRef](#)]
- [27] G. Figà Talamanca and S. Arfini, “Through the newsfeed glass: Rethinking filter bubbles and echo chambers,” *Philos. Technol.*, vol. 35, no. 1, 2022, Art. no. 20. [[CrossRef](#)]
- [28] Y. Zhao, “Book review: The atlas of AI: Power, politics, and the planetary costs of artificial intelligence,” *Glob. Media China*, 2025, Art. no. 20594364251325469. [[CrossRef](#)]
- [29] A. Vlahou et al., “Data sharing under the general data protection regulation: Time to harmonize law and research ethics?,” *Hypertension*, vol. 77, no. 4, pp. 1029–1035, 2021. [[CrossRef](#)]
- [30] L. Floridi and J. Cowls, “A unified framework of five principles for AI in society,” in *Machine Learning and the City: Applications in Architecture and Urban Design*. Hoboken, NJ, USA: Wiley, 2022, pp. 535–545. [[CrossRef](#)]
- [31] K. Patel, “Ethical reflections on data-centric AI: Balancing benefits and risks,” *SSRN*, 2024. [[CrossRef](#)]
- [32] W. Elsharif, M. Alzubaidi, and M. Agus, “Cultural bias in text-to-image models: A systematic review of bias identification, evaluation, and mitigation strategies,” *IEEE Access*, vol. 13, pp. 122636–122659, 2025. [[CrossRef](#)]
- [33] L. Apriliani, “Hypernudging in the digital era: Exploring the impact and ethical considerations of advanced behavioral interventions,” *UPY Bus. Manag. J. (UMBJ)*, vol. 4, no. 2, pp. 200–215, 2025. [[CrossRef](#)]
- [34] P. Makhijani et al., “Lime Diseases Classification Using Machine Learning and Spectrometry,” in *Information Systems for Intelligent Systems. ISBM 2024. Smart Innovation, Systems and Technologies*. Singapore: Springer, 2025, pp. 405–414. [[CrossRef](#)]
- [35] V. Braun and V. Clarke, “Conceptual and design thinking for thematic analysis,” *Qual. Psychol.*, vol. 9, no. 1, 2022, Art. no. 3. [[CrossRef](#)]
- [36] H. Tajfel, J. Turner, W. G. Austin, and S. Worchel, “An integrative theory of intergroup conflict,” *Intergroup relations: Essential readings*, pp. 94–109, 2001. [[View Online](#)]
- [37] N. Yee and J. Bailenson, “The proteus effect: The effect of transformed self-representation on behavior,” *Hum. Commun. Res.*, vol. 33, no. 3, pp. 271–290, 2007. [[CrossRef](#)]
- [38] J. Feng, H. Tan, W. Li, and M. Xie, “Conv2NeXt: Reconsidering Conv NeXt Network Design for Image Recognition,” in *Proc. 2022 Int. Conf. Comput. Artif. Intell. Technol. (CAIT)*, Nov. 4–6, 2022, pp. 53–60. [[CrossRef](#)]
- [39] G. Bernal et al., “Unraveling the dynamics of mental and visuospatial workload in virtual reality environments,” *Computers*, vol. 13, no. 10, 2024, Art. no. 246. [[CrossRef](#)]
- [40] A. B. Stanescu and D. Goanta, “Immersive tendencies and physiological responses in Virtual Reality: Enhancing education with biofeedback tools,” in *Proc. ICERI2025 Proc.*, 2025, pp. 8060–8070. [[CrossRef](#)]

- [41] K. Zhang et al., “Inclusive avatar guidelines for people with disabilities: Supporting disability representation in social virtual reality,” in *Proc. 2025 CHI Conf. Hum. Factors Comput. Syst.*, 2025, pp. 1–26. [CrossRef]
- [42] J. Cearns, “Through the looking glass: Heteronymic self and otherness in AI therapy,” *Etnográfica Rev. Cent. Rede Investig. Antropol.*, no. 29, pp. 707–710, 2025. [CrossRef]
- [43] C. R. Rogers and B. F. Skinner, “Some issues concerning the control of human behavior: A symposium,” *Science*, vol. 124, no. 3231, pp. 1057–1066, 1956. [CrossRef]
- [44] Z. Ding, Y. Huang, H. Yuan, and H. Dong, “Introduction to reinforcement learning,” *Deep Reinforcement Learning: Fundamentals, Research and Applications*. Singapore: Springer, 2020, pp. 47–123. [CrossRef]
- [45] F. Rowland, “The filter bubble: What the internet is hiding from you,” *Portal Libr. Acad.*, vol. 11, no. 4, pp. 1009–1011, 2011. [CrossRef]
- [46] A. Cavoukian, “Privacy by design in law, policy and practice a white paper for regulators,” 2011. Accessed: Mar. 15, 2025. [Online]. Available: <https://gpsbydesigncentre.com/wp-content/uploads/2022/02/312239.pdf>
- [47] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in *Proc. 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 18–24, 2022, pp. 11966–11976. [CrossRef]
- [48] T. Ma, W. Wang, and Y. Chen, “Attention is all you need: An interpretable transformer-based asset allocation approach,” *Int. Rev. Financ. Anal.*, vol. 90, 2023, Art. no. 102876. [CrossRef]
- [49] R. K. E. Bellamy et al., “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias,” *IBM J. Res. Dev.*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019. [CrossRef]
- [50] A. Rahman, “Book Review: Algorithms of oppression: How search engines reinforce racism,” *New Media Soc.*, vol. 22, no. 3, pp. 575–577, 2020. [CrossRef]
- [51] Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery,” *Queue*, vol. 16, no. 3, pp. 31–57, 2018. [CrossRef]
- [52] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017. [View Online]
- [53] R. S. Al-Mansoori, D. Al-Thani, and R. Ali, “Designing for digital wellbeing: From theory to practice a scoping review,” *Hum. Behav. Emerg. Technol.*, vol. 2023, no. 1, 2023, Art. no. 9924029. [CrossRef]
- [54] R. A. Calvo and D. Peters, *Positive Computing: Technology for Wellbeing and Human Potential*. Cambridge, MA, USA: MIT Press, 2014. [View Online]
- [55] B. Shneiderman, “Human-centered artificial intelligence: Reliable, safe & trustworthy,” *Int. J. Hum. Comput. Interact.*, vol. 36, no. 6, pp. 495–504, 2020. [CrossRef]
- [56] F. Santoni de Sio and J. Van den Hoven, “Meaningful human control over autonomous systems: A philosophical account,” *Front. Robot. AI*, vol. 5, 2018, Art. no. 323836. [CrossRef]
- [57] H. C. Triandis, “Review of culture’s consequences: International differences in work-related values,” *Hum. Organ.*, vol. 41, no. 1, pp. 86–90, 1982. [View Online]
- [58] A. C. Bluedorn, “An interview with anthropologist Edward T. Hall,” *J. Manag. Inq.*, vol. 7, no. 2, pp. 109–115, 1998. [CrossRef]

Disclaimer/Publisher’s Note: The views expressed in this article are those of the author(s) and do not necessarily reflect the views of the publisher or editors. The publisher and editors assume no responsibility for any injury or damage resulting from the use of information contained herein.